

Inaugural dissertation
for
obtaining the doctoral degree
of the
Combined Faculty of Mathematics, Engineering and Natural Sciences of the
Ruprecht - Karls - University
Heidelberg

Presented by

M. Sc. Elsa Wassmer

born in: Arras (France)

Oral examination: 15/12/2023

What makes an RNA-binding protein (RBP):
exploring the RNA-binding domains (RBDs) of
experimentally detected RBPs

Referees: Prof. Dr. Karsten Rippe

Prof. Dr. Sven Diederichs

Table of Contents

Abstract	1
Zusammenfassung.....	3
List of Figures	5
List of Tables.....	7
Abbreviations	9
1. Introduction	13
1.1. The RNA-binding proteins.....	13
1.1.1. The history of RBPs.....	13
1.1.1.1. Canonical RBPs are implicated in RNA metabolism	13
1.1.1.2. RNA can also influence non-canonical RBPs	15
1.1.2. RBPs play a role in human pathology.....	16
1.1.2.1. Implication in neurological diseases	16
1.1.2.2. The role of RBPs in cancer	17
1.2. Unraveling the RBPome	18
1.2.1. mRNA interactome capture and its variations	19
1.2.2. Techniques based on crosslinking with different capture methods	20
1.2.3. RBP screens based on other principles	22
1.2.4. Computational prediction of RBPs	25
1.2.5. New RBPs, new questions	26
1.3. How do RBPs bind RNA: the RNA-binding domains.....	29
1.3.1. The classical RBDs	29
1.3.1.1. The different types of classical RBDs	29
1.3.1.2. Classical RBDs are often arranged in a modular fashion	30

1.3.2.	The expansion of RBPomes unveiled new RBDs.....	31
1.3.3.	The role of disorder in RNA-binding.....	31
1.4.	Aim of the study.....	34
2.	Results	36
2.1.	Selection of the RNA-binding domains from the literature and the InterPro database	36
2.2.	Distribution of the selected RBDs in the species and the proteins of the RBP2GO database.....	39
2.3.	Relevance of the presence of RBDs for the selection of RBP candidates for further validation.....	43
2.4.	Selection of the RNA-related annotations from the literature and the InterPro database	48
2.5.	Rfam IDs also help to discriminate strong RBP candidates.....	51
2.6.	Distribution of disordered regions in the proteins of the RBP2GO database	56
2.7.	Prediction of new RBP candidates	64
2.8.	Prediction of new RBD candidates in proteins detected as RBPs but lacking known RBDs.....	67
2.9.	Validation of the new RBDs using published data on RNA-binding peptides.....	71
2.10.	Establishment of a new RBP2GO composite score	73
2.11.	Implementation of the new data in the RBP2GO database.....	76
3.	Discussion	80
4.	Materials.....	89

5.	Methods	92
5.1.	Compilation of a list of RNA-binding domain candidates.....	92
5.2.	Selection of the RNA-binding domains	92
5.3.	Protein expression levels in HeLa cells.....	93
5.4.	Retrieval of the InterPro domain coordinates and compilation of the number and content fraction of RBDs per protein	94
5.5.	Construction of a list of RNA-related family IDs and selection of the IDs enriched in RBPs.	94
5.6.	Retrieval of the MobiDB-lite data for disordered regions	95
5.7.	GO enrichment analysis for non-RBPs containing an RBD	95
5.8.	Identification of new RBDs using the human proteins	96
5.9.	Validation of the newly discovered RBDs using published lists of RNA-binding peptides.....	97
5.10.	Computation of the new RBP2GO composite score.....	97
5.11.	Selection of a list of high-confidence RBPs in Hs	99
5.12.	Update of the RBP2GO database	99
6.	Acknowledgements	101
7.	Appendix	102
8.	References	128

Abstract

RNA-binding proteins, or RBPs, are responsible for the regulation of RNA fate from transcription to decay. In the past 20 years, their implications in human pathology have been highlighted, especially in hereditary neurodegenerative diseases but also in the development of cancer. Furthermore, in recent studies, the emergence of long non-coding RNAs (lncRNAs) questioned the established dogma and showed that RNAs can also influence the fate of RBPs, through the regulation of their localization, interactions, or activation. Therefore, it became urgent to systematically detect the proteins able to bind to RNA, and several different techniques have been developed in the past ten years to address this challenge. However, the accumulation of published lists of RBPs toughened the access to comprehensive data. Subsequently, the RBP2GO database was created. This database compiles all of the proteome-wide screens available in the literature, and facilitate the access of scientists to this ever-growing mass of information.

The multiplicity of proteome-wide RBP screens also brings into question the specificity of the published data. Indeed, the number of RBPs in human has quickly risen up to a third of the total proteome, and little overlap can be found between the different datasets. Furthermore, most of the techniques employed do not allow the identification of the precise part of the protein which is binding RNA. As a result, no information on RNA-binding features was available for the RBPs of the RBP2GO database. Hence, I exploited the data available in this database to study the presence of RNA-binding domains in experimentally detected RBPs. The aim of this analysis was to determine if these domains could be used to better segregate relevant RBP candidates. I first compiled a list of RNA-binding domains (RBDs), and selected them based on their enrichment in RBPs to then dissect their repartition on the proteins of the database. The same was done for RNA-related family IDs (Rfam IDs), as well as disordered regions. This bioinformatic analysis showed that RBDs and Rfam IDs are strong indicators of the RNA-binding potential of proteins. However, the presence of disorder did not appear as important, and a higher proportion of disorder was observed in the proteins already exhibiting an RBD. This gained knowledge was used to predict new RBP candidates. The RBPs with no RBD were also studied, and 15 new RBDs were predicted and subsequently validated using RNA-binding peptides from mass spectrometry data. Finally, a new score, called the RBP2GO composite score, was created as a single metric assembling both experimental RBPome data and the

presence of RBDs or Rfam IDs. This score was used to then compile a list of high-confidence human RNA-binding proteins. All of this newly acquired information was integrated into the RBP2GO database (<https://RBP2GO.dkfz.de>), to provide an easy access to future users.

Zusammenfassung

RNA-bindende Proteine (RBPs) sind regulieren den Lebenszyklus einer RNA von der Transkription bis zum Abbau. In den letzten 20 Jahren wurde ihre Rolle in verschiedenen Krankheiten charakterisiert, insbesondere für erbliche neurodegenerative Erkrankungen, aber auch für die Entstehung von Krebs. Darüber hinaus hat die Entdeckung von langen nicht-kodierenden RNAs (lncRNAs) das etablierte Dogma der Molekularbiologie in Frage gestellt und gezeigt, dass RNAs auch die Funktion von RBPs beeinflussen können, indem sie deren Lokalisierung, Interaktionen oder Aktivierung regulieren können. Daher wurden in den letzten 10 Jahren verschiedene Techniken entwickelt, um systematisch alle Proteine zu identifizieren, die an RNA binden können. Doch die steigende Anzahl von Listen von RBPs erschwerte den Zugang zu den umfassenden Datensätzen aus verschiedenen Quellen. Daher wurde die RBP2GO-Datenbank entwickelt, um alle in der Literatur verfügbaren proteomweiten Screens zusammenzustellen und anderen Wissenschaftlern den Zugang zu diesen Informationen zu erleichtern.

Die Auswertung der großen Anzahl an proteomweiten RBP-Screens stellt auch deren Spezifität in Frage. In der Tat ist die Zahl der RBPs beim Menschen schnell bis auf ein Drittel des gesamten Proteoms angestiegen, aber die Datensätze überlappen sich kaum. Darüber hinaus erlauben die meisten der verwendeten Techniken nicht die Identifizierung des genauen Teils des Proteins, der RNA bindet, so dass für die RBPs der RBP2GO-Datenbank keine Informationen über RNA-bindende Merkmale verfügbar sind. Daher habe ich die in dieser Datenbank verfügbaren Daten genutzt, um das Vorhandensein von RNA-bindenden Domänen in experimentell nachgewiesenen RBPs zu untersuchen. Ziel dieser Analyse war es, festzustellen, ob diese Domänen genutzt werden können, um relevante RBP-Kandidaten besser auszusortieren. Ich habe zunächst eine Liste von RNA-bindenden Domänen (RBDs) zusammengestellt und daraus RBDs basierend auf ihrer Anreicherung in RBPs ausgewählt, um dann ihre Verteilung in den Proteinen der Datenbank zu untersuchen. Das Gleiche wurde für RNA-verwandte Familien-IDs (Rfam IDs) sowie für intrinsisch ungeordnete Domänen durchgeführt. Diese bioinformatische Analyse zeigte, dass RBDs und Rfam IDs starke Indikatoren für das RNA-Bindungspotenzial von Proteinen sind. Das Vorhandensein von ungeordneten Domänen schien jedoch keine eigenständige Rolle zu spielen, und ein höherer Anteil an Unordnung wurde in den Proteinen beobachtet, die bereits eine RBD aufwiesen. Diese gewonnenen Erkenntnisse wurden einerseits genutzt, um neue RBP-Kandidaten

vorherzusagen. Andererseits wurden auch die RBPs ohne RBD untersucht, und es wurden 15 neue RBDs vorhergesagt und anschließend mit RNA-bindenden Peptiden aus massenspektrometrischen Daten validiert. Schließlich wurde ein neuer Score, der so genannte RBP2GO Composite Score, als eine einzige Metrik erstellt, die sowohl experimentelle Daten aus proteomweiten RBP-Screens, als auch das Vorhandensein von RBDs und Rfam-IDs berücksichtigt. Dieser Score wurde dann verwendet, um eine Liste von menschlichen RNA-bindenden Proteinen mit hoher Konfidenz zusammenzustellen. Alle diese neu gewonnenen Informationen wurden in die RBP2GO-Datenbank (<https://RBP2GO.dkfz.de>) integriert, um künftigen Nutzern einen einfachen Zugang zu ermöglichen.

List of Figures

Figure 1: Schematic representation of the cross-linking based RBP screens.....	21
Figure 2: Schematic representation of the other RBP screens	23
Figure 3: Upset plots showing the number of proteins detected as RBPs in one or more datasets in <i>Homo sapiens</i> (Hs), <i>Mus musculus</i> (Mm), <i>Saccharomyces cerevisiae</i> (Sc) and <i>Drosophila melanogaster</i> (Dm).....	28
Figure 4: Graphical abstract summarizing the study presented in this thesis	35
Figure 5: Selection process of the RNA-binding domains.....	37
Figure 6: Scatterplot displaying the number of datasets compiled in the RBP2GO database for each species.....	38
Figure 7: Analysis of the distribution of RBDs in the species of the RBP2GO database.....	40
Figure 8: Distribution of the RBDs in the RBPs and non-RBPs of the RBP2GO database and correlation with their expression.....	42
Figure 9: Influence of the presence of RBDs on the RBP2GO score.....	44
Figure 10: Proportion of RBD-containing proteins (in %) for each unit of RBP2GO score.....	45
Figure 11: Influence of the number of RBDs on the RBP2GO score.....	46
Figure 12: Selection process of the RNA-related family IDs (Rfam IDs).....	48
Figure 13: Repartition of the Rfam IDs in the species and the proteins of the RBP2GO database.....	50
Figure 14: Influence of the presence of Rfam IDs on the RBP2GO score.....	52
Figure 15: Proportion of Rfam ID-associated proteins (in %) for each unit of RBP2GO score	53
Figure 16: Influence of the presence of Rfam IDs in combination with RBDs on the RBP2GO score.....	55
Figure 17: Repartition of the disordered regions in the proteins of the RBP2GO database.....	57
Figure 18: Repartition of the polyampholyte IDRs in the proteins of the RBP2GO database.....	59
Figure 19: Repartition of the coiled-coil regions in Hs, Mm, Sc and Dm.....	60
Figure 20: Influence of the presence of an IDR on the RBP2GO score.....	62

Figure 21: Correlation between the presence of disordered regions and the RBP2GO score.....	63
Figure 22: Discovery of new RBP candidates using the RBD-content of the proteins.....	65
Figure 23: Discovery of new RBD candidates in RBPs with no RBD.....	68
Figure 24: Validation of the new RBD candidates in the most studied species of the RBP2GO database.....	70
Figure 25: Validation of the new RBD candidates using experimentally identified RNA-binding peptides.....	72
Figure 26: Development of the new RBP2GO composite score.....	74
Figure 27: Update of the RBP2GO database to integrate the new data.....	77
Figure 28: Information about the RBDs and the Rfam IDs integrated in the RBP2GO database.....	78
Figure 29: Integration of the generated data in the RBP2GO database.....	79
Figure 30: Repartition of human RBPs with no RBD and detected only once in the human datasets.....	87

List of Tables

Table 1: List of selected RBDs.....	on CD
Table 2: New RBP candidates validated by a literature search.....	66
Table 3: Attribution rules of quality factors to RBDs and Rfam IDs.....	73
Table 4: List of the R packages used for the analysis presented in this thesis.....	89
Table 5: List of the R packages used to generate the figures presented in this thesis.....	90
Table 6: List of the R packages used for the RBP2GO database.....	91
Table 7: List of selected RNA-related family IDs.....	on CD
Table 8: List of all proteins of the RBP2GO database with information on RBDs, RNA-related family IDs and disorder.....	on CD
Table 9: Results of the GO term enrichment analysis for the biological process terms, limited to terms enriched in at least five species.....	102
Table 10: Results of the GO term enrichment analysis for the molecular function terms, limited to the terms enriched in at least five species.....	114
Table 11: List of newly discovered RBDs and their enrichment.....	126
Table 12: List of the RBDs and Rfam IDs with their quality factors.....	on CD

Abbreviations

Abbreviation	Full form
%	Percent
4SU	4-thiouridine
6SG	6-thioguanosine
AAindex	amino acid index
ADAR2	adenosine deaminase RNA specific 2
ALS	amyotrophic lateral sclerosis
Arm	Armadillo
ARMC	Armadillo-repeat containing
ASO	antisense oligonucleotide
At	<i>Arabidopsis thaliana</i>
BLAST	basic local alignment search tool
BPS	branch point sequence
C/EBP β	CCAAT Enhancer Binding Protein Beta
CAID	critical assessment of protein disorder
CAPRI	crosslinked and adjacent peptides-based RNA-binding domain identification
CARIC	click chemistry-assisted RNA interactome capture
CASP10	critical assessment of methods for protein structure prediction
cCL	conventional UV crosslinking
CDK	cyclin-dependent kinase
Ce	<i>Caenorhabditis elegans</i>
CLIP	crosslinking immunoprecipitation
coREST	REST corepressor
DICE	differentiation control element
Dm	<i>Drosophila melanogaster</i>
DM1/2	myotonic dystrophy type 1/2
DMPK	myotonic dystrophy protein kinase
DNA	deoxyribonucleic acid
Dr	<i>Danio rerio</i>
dsDNA	double-stranded DNA
DSEF-1	downstream-element factor 1
dsRBD	double-stranded RNA binding domain
dsRNA	double-stranded RNA
EBI	european bioinformatics institute
Ec	<i>Escherichia coli</i>
EJC	exon junction complex
ELAV1	ELAV-like protein 1
eRIC	enhanced RIC
ERK	extracellular signal-regulated kinase

EU	ethynyl-uridine
FMRP	fragile X messenger ribonucleoprotein
FSX	fragile X syndrome
FUS	fused in sarcoma
GAPDH	glyceraldehyde-3-phosphate dehydrogenase
Gas5	growth arrest specific 5
GO	gene ontology
GR	glucocorticoid receptor
HMGB	high mobility group box
hnRNA	heterogeneous nuclear RNA
hnRNP	heterogeneous nuclear RNP
HOTAIR	HOX transcript antisense RNA
HOXD	homeobox D
Hs	<i>Homo sapiens</i>
HSP	heat shock protein
HuR	Hu-antigen R
IDR	intrinsically disordered region
IRE	iron response element
KH domain	K-homology domain
Ld	<i>Leishmania donovani</i>
Lm	<i>Leishmania mexicana</i>
LNA	locked nucleic acid
LOX	15-lipoxygenase
LSD1	lysine specific demethylase 1
magoh	Mago homologue
MALAT1	metastasis-associated lung adenocarcinoma transcript 1
MBNL	muscle blind like protein family
miRNA	microRNA
Mm	<i>Mus musculus</i>
Mre11	meiotic recombination 11 homolog 1
mRNA	messenger RNA
mRNP	messenger RNP
NEAT2	nuclear-enriched transcript 2
NKAP	NF-kappa-B-activating protein
NMD	non-sense mediated decay
nt	nucleotide
PAR-CLIP	photoactivable-ribonucleoside-enhanced crosslinking immunoprecipitation
PAZ	Piwi, Argonaut and Zwillie
pCLAP	peptide crosslinking and affinity purification
PDB	protein data bank
Pf	<i>Plasmodium falciparum</i>
PPI	protein-protein interaction

pre-mRNA	pre-messenger RNA
PSI-BLAST	position-specific iterative BLAST
PTM	post-translational modification
QUA2	Quaking homology 2
RBD	RNA-binding domain
RBDmap	RNA-binding domain mapping
RBP	RNA-binding protein
RBR-ID	RNA-binding region identification
REST	RE1 silencing transcription factor
Rfam ID	RNA-related family ID
RIC	RNA interactome capture
RICK	RNA interactome using click chemistry
RNA	Ribonucleic acid
RNP	ribonucleoparticle
RNPS1	RNA-Binding Protein With Serine-Rich Domain 1
RRM	RNA recognition motif
Sc	<i>Saccharomyces cerevisiae</i>
SDAD1	SDA1 domain-containing protein 1
SDS-PAGE	sodium dodecyl sulfate polyacrylamide gel electrophoresis
serIC	serial interactome capture
Sp-1	Specificity-protein 1
Srm160	SR-related nuclear matrix protein of 160 kDa
SRSF1	serine and arginine-rich splicing factor 1
ST	<i>Salmonella Typhimurium</i>
TAP	Transporter
Tb	<i>Trypanosoma brucei</i>
TDP43	TAR DNA-binding protein 43
TFIIIA	Transcription factor III A
TRAPP	total RNA-associated protein purification
Upf3	up-frameshift suppressor 3
UTR	untranslated region
UV	ultra-violet
WDR43	WD40 repeat containing protein 43
ZNF9	zinc finger protein 9

1. Introduction

1.1. The RNA-binding proteins

RNA-binding proteins, or RBPs, are the main focus of this thesis. Therefore, this chapter will introduce the historical discovery of RBPs, as well as the most recent advances in this field.

1.1.1. The history of RBPs

1.1.1.1. Canonical RBPs are implicated in RNA metabolism

In 1958, Francis Crick proposed what is now known as the central dogma of molecular biology: the genetic information is encoded in the DNA present in the nucleus of cells, and is transported outside of the nucleus via messenger RNAs, or mRNAs, that are later translated into proteins (1, 2). Furthermore, some groups studied RNase-sensitive granules in the 1950s and 1960s, exposing that the nuclear RNAs, called hnRNAs (heterogeneous nuclear RNAs) are always coated by proteins (3–5). Subsequent research revealed that these RNA-binding proteins associate with the hnRNA molecule during its transcription, and participate in all stages of RNA metabolism, from its transcription to its translation and degradation (6, 7). These proteins are called hnRNP proteins, for heterogeneous nuclear ribonucleo-particles proteins. At the same time, proteins associated with cytoplasmic mRNAs were discovered, forming complexes called mRNPs, or messenger ribonucleo-particles (8, 9). Those two classes of proteins were then distinguished by their subcellular localization and the fact that they never co-precipitated with each other on the same target RNA (7). However, little was known about hnRNPs, mRNPs, and their exact function in the processing of mRNAs.

In the 1980s, experiments combined the isolation of nuclei from human cells, and the separation of the different proteins contained in hnRNPs on a sucrose gradient. Subsequent RNase treatment was used to free the proteins from their bound RNA, permitting their isolation and their characterization. hnRNP proteins were then named with letters, in the order of their discovery (9, 10). Their study revealed that these proteins are implicated in all steps of mRNA processing, starting with transcription. For example, the hnRNPK protein has been found to bind CT elements upstream of the c-myc gene. When interacting with the Sp-1 (Specificity-protein 1) transcription factor, it stimulates the transcription of c-myc (11). But it can also interact with the C/EBP β (CCAAT enhancer binding protein beta) transcription factor and inhibit its trans-activation of the agp gene (12). hnRNPA1 expression has also been shown to

modulate the selection of 5' splice sites in several pre-mRNAs, demonstrating that hnRNP proteins are also involved in the regulation of splicing (13, 14). Furthermore, DSEF-1 (downstream-element factor 1), a protein of the hnRNPH family, can bind the AAUAAA sequence in the 3' UTR of multiple mRNAs *in vitro*, and stimulate their cleavage and polyadenylation (15). hnRNP proteins are thus involved in mRNA processing, and subsequently influence their stability.

Further experiments showed that hnRNP proteins functions extend beyond the nucleus. For example, hnRNPA1 is also involved in the export of mRNAs outside of the nucleus, thanks to its M9 sequence. This pathway is completely independent from the importin α/β pathway, and requires the transportin protein. hnRNPA1 shuttles between the nucleus and the cytoplasm, contradicting the spatial restriction of hnRNP proteins to the nucleus. Moreover, this pathway is specific to mRNAs and its perturbation does not influence the localization of other types of RNAs (16, 17). In addition, the hnRNPK/E1 protein accumulates in the cytoplasm upon phosphorylation by the ERK (extracellular signal-regulated kinase) in response to growth factors. It then binds to the differentiation control element (DICE) of the 15-lipoxygenase (LOX) mRNA in the 3' UTR. There, the 40S ribosomal subunit is able to bind the mRNA, but the 60S subunit cannot be recruited, and the translation of the LOX protein is inhibited (18, 19). Hence it became clear, at the beginning of the 21st century, that hnRNP proteins can bind RNA and influence its fate from transcription until translation. Additionally, they are not spatially restricted to the nucleus. Their historical separation with the mRNP proteins thus appears outdated.

Furthermore, several proteins exhibit a role in different steps of mRNA metabolism, which allows for the coupling between these different steps (7). A relevant illustration of this concept is the Exon Junction Complex, or EJC. It was discovered at the beginning of the 2000s thanks to the development of a new technique of cross-linking that made the retrieval of proteins bound to a specific RNA possible. This technique in turn made the comparison of the mRNP composition between unspliced and spliced mRNAs possible. The EJC is comprised of 6 proteins, Srm160 (SR-related nuclear matrix protein of 160 kDa), RNPS1 (RNA-Binding Protein With Serine-Rich Domain 1), Aly/REF, y14, magoh (Mago homolog) and Upf3 (up-frameshift suppressor 3), forming a complex sitting at the junction of two exons which is deposited during splicing (7, 20). This complex is indispensable for the efficient export of the mRNA in the cytoplasm. It was already known that splicing is required for the export of cellular

mRNAs, but the responsible proteins had not yet been isolated (21). Aly, a member of the EJC, interacts with the TAP (Transporter) protein and facilitates the export of spliced mRNAs. In this way, the TAP-Aly interaction couples the splicing of mRNAs with their export (7, 22). Proteins of the EJC are also important for the NMD, or non-sense mediated decay pathway. If the mRNA possesses a stop codon 50 to 55 nt (nucleotides) upstream of the last exon, this pathway is triggered during translation, and the RNA is degraded (23). Normally, all the EJCs deposited at exon-exon junctions are removed from the mRNA during the first round of translation. But if a stop codon is present before an exon-exon junction, some EJCs will remain and signal the mRNA to the NMD. The proteins of the EJC are thus carrying the information about the position of former introns, and are able to couple splicing and the NMD, two physically separated processes (7). This shows that RNA-binding proteins carry messages and are able to coordinate the different steps of RNA processing.

1.1.1.2. RNA can also influence non-canonical RBPs

All of the RNA-binding proteins implicated in RNA metabolism are today referred to as canonical RBPs, in contrast to the other RBPs which are not implicated in these pathways (24). Canonical RBPs control the fate of their target RNA, at any step between its transcription and its degradation.

The discovery of long non-coding RNAs (lncRNAs) in the 2000s, and their implication in transcription and chromatin regulation, highlighted the existence of non-canonical RBPs (25). Thought for a long time to only represent transcriptional noise, lncRNAs are RNAs longer than 200 nt which do not have a coding potential. They can be transcribed from coding genes, as well as introns or intergenic regions, in any direction (25). These RNAs are able to recruit large chromatin modifying complexes, and target them to a given position in the genome. For instance, HOTAIR (HOX transcript antisense RNA) can bind both the polycomb repressive complex 2 (PRC2) and the LDS1/coREST/REST (lysine specific demethylase / REST corepressor / RE1 silencing transcription factor) complex. Both of these complexes modify marks of the histone H3. PRC2 methylates the lysine 27 and coREST/REST demethylates the lysine 4 (26–28). HOTAIR targets both of these complexes to the HOXD (homeobox D) gene locus, compacting the chromatin and silencing this gene (29). Other lncRNAs can act as a decoy (25), such as growth arrest specific 5, or Gas5. Gas5 binds to the glucocorticoid receptor, or GR, and titrates it away from its binding sites in the genome. Subsequently, the GR cannot

activate the transcription of its target genes (30). lncRNAs can also ensure other functions, such as acting as a scaffold for the assembly of protein complexes, or as an enhancer (25).

Recent work on ribonucleoprotein (RNP) granules also underscores the influence of RNA on protein function. Some proteins, over a certain threshold concentration, are able to phase separate and form liquid to solid membrane-less condensates within the cytoplasm or the nucleus, such as P granules, Cajal bodies, the nucleolus or paraspeckles (31). Some of these proteins are also able to bind to RNA, such as Fused in Sarcoma (Fus), a prion-like protein (32). This protein binds to the C-terminal domain of the RNA polymerase II, and can affect transcription, but is also found to phase separate in the cytoplasm thanks to its intrinsically disordered region or IDR (33). RNA has been shown to modulate the behavior of Fus: a high RNA/protein ratio prevents the formation of granules, while a low ratio promotes phase separation, explaining the different behaviors of Fus in the nucleus and in the cytoplasm (34). Moreover, RNA can also regulate the viscosity of RNPs. In *C. elegans*, the LAF-1 protein is part of the P granules, and can phase separate *in vitro*. The viscosity of the droplets can be modified via the concentrations in salt and RNA, higher RNA concentrations enhancing their fluidity (35).

These few selected discoveries highlight that RNAs can bind non-canonical RBPs, which are not implicated in RNA metabolism, and influence their function, localization and/or interactions with other proteins. This adds to the central concept that the fate of RBPs is influenced by RNA, and highlights the importance to detect non-canonical RBPs whose function might be regulated by RNA.

1.1.2. RBPs play a role in human pathology

1.1.2.1. Implication in neurological diseases

Neurons are highly polarized and organized cells and have been shown to regulate mRNA metabolism at the subcellular level to ensure the right localization of the produced proteins (36). As a result, a lot of neurological and neuromuscular diseases find their cause in the dysregulation of RBPs, especially prion-like proteins such as Fus or TDP43 (TAR DNA-binding protein 43) (37, 38). Three mechanisms can be implicated: the reduced expression level of an RBP, its increased propensity to form aggregates, or its sequestration by abnormal RNAs (38).

The fragile X syndrome (FXS) is a prime example of a neurological disease caused by the mutation of an RBP. FXS is an inherited neurodevelopmental disorder with autistic symptoms, present in 1 in 4,000 males and 1 in 8,000 females. Its cause is the presence of more than 200 CGG microsatellite repeats in the 5' UTR of the FMR1 gene, coding for the FMRP (Fragile X messenger ribonucleoprotein) protein, against 6 to 54 in healthy individuals (39, 40). These repeats are methylated, resulting in the silencing of the FMR1 gene (39–41). The FMRP protein is particularly expressed in neurons, where it binds to mRNAs and ensures their local translation at the dendrites. Hence, its down-regulation results in immature dendritic spines in the brains of FXS patients (42).

Amyotrophic lateral sclerosis, or ALS, is a rare neurodegenerative condition, characterized by the death of motor neurons in the brain and spinal cord, leading to progressive muscle weakness and ultimately fatal paralysis (43). This disease has been linked to mutations in the gene coding for the TDP43 (TAR DNA-binding protein 43) (44). This protein is normally present in the nucleus of cells, where it regulates RNA splicing. In 97% of ALS cases, aggregates of ubiquitinated and heavily phosphorylated TDP43 are found in the cytoplasm of neurons, causing its depletion from the nucleus (45).

Myotonic dystrophies types 1 and 2 (DM1 and DM2) are autosomal dominant syndromes characterized by muscle weakness and myotonia (46). DM1 is characterized by the presence of 50 to 3,500 CTG repeats in the 3' UTR of the DMPK (myotonic dystrophy protein kinase) gene (47), while DM2 is caused by 75 to 11,000 CCTG repeats in the first intron of the ZNF9 (zinc finger protein 9) gene (48). Both repeat expansions lead to the expression of transcripts which sequester RBPs from the muscle blind like protein family (MBNL1 to 3) (49, 50). The MBNL proteins are involved in the regulation of splicing, and their disruption was shown to replicate DM symptoms in mice, while their overexpression can rescue the DM phenotype (38).

1.1.2.2. The role of RBPs in cancer

Considering that RBPs are implicated in gene expression through their regulation of RNAs, it is not surprising that several RBPs have been found to be important in the development of cancer. They can participate in all hallmarks of cancer, generally several at the same time, and are thus very interesting targets for therapy (51, 52). They are commonly upregulated in cancer

cells, with very few mutations observed but rather copy-number variations (CNVs) with gene amplifications leading to an over-expression of these proteins (52).

A prime example of the role of RBPs in cancer is the Hu-antigen R (HuR) protein, also called ELAV1 (ELAV-like protein 1). This protein is overexpressed in a number of different cancers, including brain, breast, cervical, colon, and lung cancers (52). HuR enhances the stability of different mRNAs which encode for pro-proliferative genes such as cyclins, via the binding of AU-rich elements (AREs) in their 3' UTR (53–55). It also enhances the translation of anti-apoptotic proteins, participating in the survival of cancer cells (56). As a result, HuR can be considered as a “master” of gene expression in cancer cells and participates in almost every hallmark of cancer (57). Its overexpression has also been linked to tumor aggressiveness and poor outcomes in several tumor types (58). Its knockout in different cancer models leads to an attenuated tumor growth (59, 60), and HuR is quickly activated by a number of different therapies, highlighting its role in therapy resistance (61). Therefore, targeting HuR appears as a promising strategy to sensitize tumor cells to treatment (62). Small molecules inhibiting HuR and strategies to prevent its interaction with its target RNAs are currently in development, with the hope to bring a treatment to sensitize cancer cells to chemotherapeutics in the clinic (57).

Several lncRNAs are also overexpressed in cancer, disrupting the localization or interactions of their bound RBPs (63). For example, the metastasis-associated lung adenocarcinoma transcript 1, or MALAT1, also known as nuclear-enriched transcript 2 (NEAT2) is a lncRNA discovered as a prognostic marker for lung cancer metastasis (64). It is overexpressed in lung cancer, but also in other cancer types such as breast cancer and liver cancer (65). The down-regulation of MALAT1 using antisense oligonucleotides (ASOs) prevents *in vivo* lung cancer metastasis, and results in slower growth, cell differentiation, and reduction in the rate of metastasis of a breast cancer mouse model (64, 66). In this regard, MALAT1 represents a promising therapeutic target. But its knock-down results in different effects in different cell lines, underscoring the importance to identify the impacted RBPs and their function (67).

1.2. Unraveling the RBPome

In view of the importance of RBPs, in genetic neuronal diseases but also in the development of cancer, several techniques have been developed in the last decade to detect RBPs in cells in

a high-throughput manner. The present chapter will present the different experimental and computational techniques that have been developed, as well as their advantages and limitations.

1.2.1. mRNA interactome capture and its variations

In 2012, the first proteome-wide studies on RBPs were published, using a method that was later called RIC, for RNA interactome capture (68, 69). This method consists in crosslinking the RNA with its bound proteins using 254 nm UV light (conventional UV crosslinking or cCL). Photoactivable nucleotide 4-thiouridine (4SU) or 6-thioguanosine (6SG), which are metabolically incorporated into transcribed RNAs and crosslinked to proteins using 365 nm UV light (photoactivable-ribonucleoside-enhanced crosslinking or PAR-CLIP) can also be used (70, 71). Both methods require a direct contact between the RNA and the protein for their crosslink to happen. Castello et al. applied both methods in parallel on HeLa cells, followed by a capture of the mRNAs using oligo-dT magnetic beads, an RNase treatment to release the proteins, and their identification via mass spectrometry (72) (Figure 1). A control experiment was used to remove any background. This resulted in the identification of 1316 RBPs in HeLa cells, among which known canonical RBPs such as the proteins of the exon junction complex, but also 315 unknown RBPs (68). Another proteome-wide RBP screen using only the PAR-CLIP crosslinking method was published the same year, yielding 800 RBPs in HEK293 cells, among which 15% were not known or predictable by computing methods (69).

Other methods based on the same principle of UV crosslinking and mRNA capture were subsequently developed, to try to circumvent the limitations of the RIC method. To improve the specific isolation of RBPs, serIC (serial interactome capture) used serial mRNA capture with oligo-dT beads. This study also introduced intermittent enzyme digestion steps to allow the detection of dual DNA-RNA binders (73). An enhanced RNA-interactome capture (eRIC) technique was also developed, using a locked nucleic acid (LNA)-modified probe instead of the oligo-dT beads to improve the capture of RNA-protein complexes and support more stringent washing conditions (74).

Further improvements on the method aimed at a better resolution of the portion of the RBP sequence which is binding to the crosslinked RNA. RBDmap (RNA-binding domain mapping) was developed from the RIC technique, and aimed at the detection of the protein sequence bound by RNAs. To do so, the release of proteins from the crosslinked RNA was modified,

using a protease digestion first to separate the mRNA-bound peptides from the rest of the protein, and then a trypsin and RNase treatment to release the peptides from the RNAs and prepare them for mass spectrometry. This method detected canonical as well as new RNA-binding domains (RBDs) (75). Another method, peptide crosslinking and affinity purification (pCLAP), used trypsin treatment of the crosslinked proteins before their isolation with oligo-dT beads to detect RNA-binding regions with a higher sensitivity than RBDmap and proposed a simplified analysis of the collected data (76). This study identified RNA-bound peptides already identified by RBDmap, in addition to several others, but only used peptides adjacent to the cross-link site. On the contrary, CAPRI (crosslinked and adjacent peptides-based RNA-binding domain identification) detected both the crosslinked peptides and the adjacent peptides to better determine the RBDs. Finally, RBR-ID (RNA-binding region identification) was based on the PAR-CLIP method. This permitted the simultaneous mass spectrometry analysis of all peptides, and the identification of RNA-binding peptides thanks to the presence of 4SU in the crosslinked RNAs (77).

1.2.2. Techniques based on crosslinking with different capture methods

One major limitation of the RIC method was the use of oligo-dT beads to isolate RNA-protein complexes. This restrained the detected RBPs to proteins binding poly-adenylated RNAs, and completely ignored the proteins binding to most non-coding RNAs. To circumvent this problem, an approach termed RICK (RNA interactome using click chemistry) took advantage of the click chemistry to label RNAs with biotin. Shortly, the cells were incubated with 5-ethynyluridine (EU), which is incorporated into the RNAs (Figure 1). They were then crosslinked with 254nm UV light and lysed. The lysate was submitted to a click reaction, which binds biotin to the EU incorporated in the RNAs (78). Finally, the RNA-protein complexes were isolated using streptavidin-coated beads. This method allowed for the capture of all RNAs that incorporated EU, without any selection for poly-adenylated RNAs (79). The same principle was used in CARIC (click chemistry-assisted RNA interactome capture), combined with the PAR-CLIP crosslinking (80).

The TRAPP (total RNA-associated protein purification) technique, on the other hand, was based on a one step lysis and isolation of the RNA-protein complexes. The cells were crosslinked, using cCL or PAR-CLIP, and lysed in a buffer containing guanidine thiocyanate and phenol. The lysate was cleared by centrifugation and the RNA-protein complexes were isolated on silica beads, that were then eluted to recover the complexes (Figure 1). The RNAs

were digested and the proteins identified by tandem mass spectrometry. The advantage of this technique, like for RICK and CARIC, was the absence of selection of the isolated RNAs, yielding a comprehensive RBPome (81).

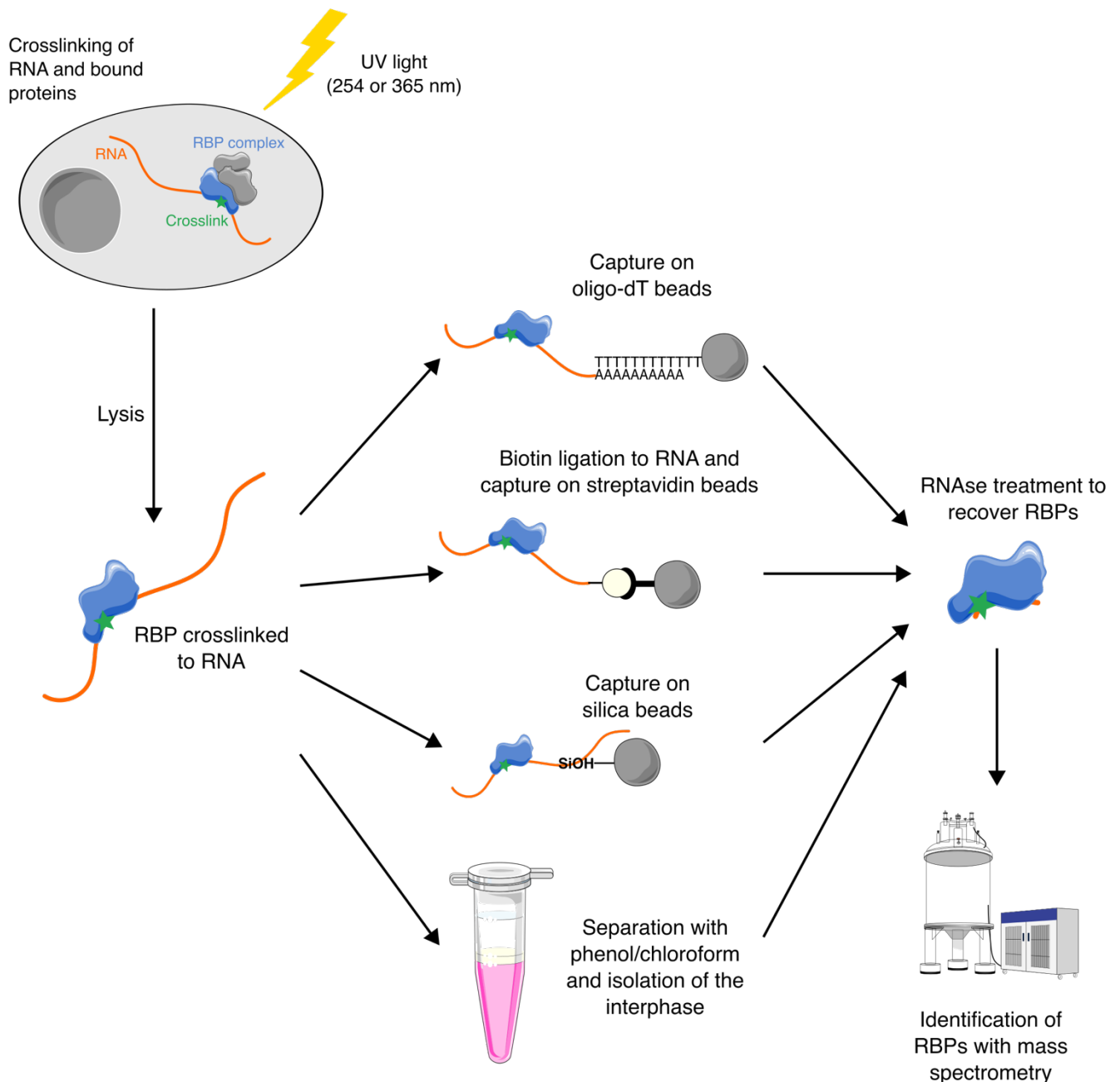


Figure 1: Schematic representation of the cross-linking based RBP-screens

Finally, some groups took advantage of the specific physico-chemical properties of the crosslinked RNA-protein complexes to isolate them. In a cell lysate, RNA can be separated

from the proteins and the DNA through an acidic phenol-chloroform extraction, with the eventual addition of guanidium thiocyanate to preserve its integrity. The solution of acidic phenol-chloroform is added to the lysate, then mixed, and the phases are separated by centrifugation. The aqueous phase then contains the RNA, while the organic phase contains the proteins and the DNA, and the interphase contains the rest of the cellular components (82).

Following this principle, a complex formed by crosslinked RNA and protein should separate in the interphase. Several groups have used this principle to isolate RBPs: the cells were irradiated with 254 nm UV light, and then lysed with a solution of phenol-chloroform or phenol-toluol. The interphase was subsequently isolated to recover the RNP complexes (Figure 1). The orthogonal organic phase separation (OOPS) used a first AGPC (guanidium-thiocyanate phenol-chloroform) extraction to isolate the RNA-protein complexes in the interphase, followed by a protease digestion and subsequent separation of RNAs and proteins in a second AGPC extraction (83). XRNAX (protein-crosslinked RNA extraction) used only one AGPC extraction, followed by washing steps of the interphase to get rid of lipids, free DNA, RNA and proteins (84). Finally, PTex (phenol toluol extraction) was based on two phenol-toluol extractions of the cell lysate. The first one at a pH of 7 allowed to retrieve RNA, proteins and crosslinked complexes in the aqueous phase, separating them from lipids and DNA. The second one, at a pH below 5, isolated the crosslinked complexes in the interphase (85). These methods have the advantage to isolate all RNA-protein complexes, without any selection bias. However, they also isolate other DNA, RNA, proteins and lipids, which then need to be washed from the interphase before the identification of crosslinked proteins and RNA, causing a higher background than for the previously presented techniques.

1.2.3. RBP screens based on other principles

Crosslinking RNA and proteins in a cell to then identify the bound RBPs has the advantage to covalently bind the two interaction partners, but the main hurdle that has to be faced is the correct isolation of these crosslinked complexes. One can target a specific subset of RNAs, using oligo-dT beads for example, but this will be limited by a selection bias. On the other hand, the attempts to isolate all the complexes at once are faced with the isolation of other molecules. Furthermore, the crosslinking step is not without bias, and unspecific interactions can be caught while functionally important ones remain undetected. To face these issues, other experimental setups have been developed.

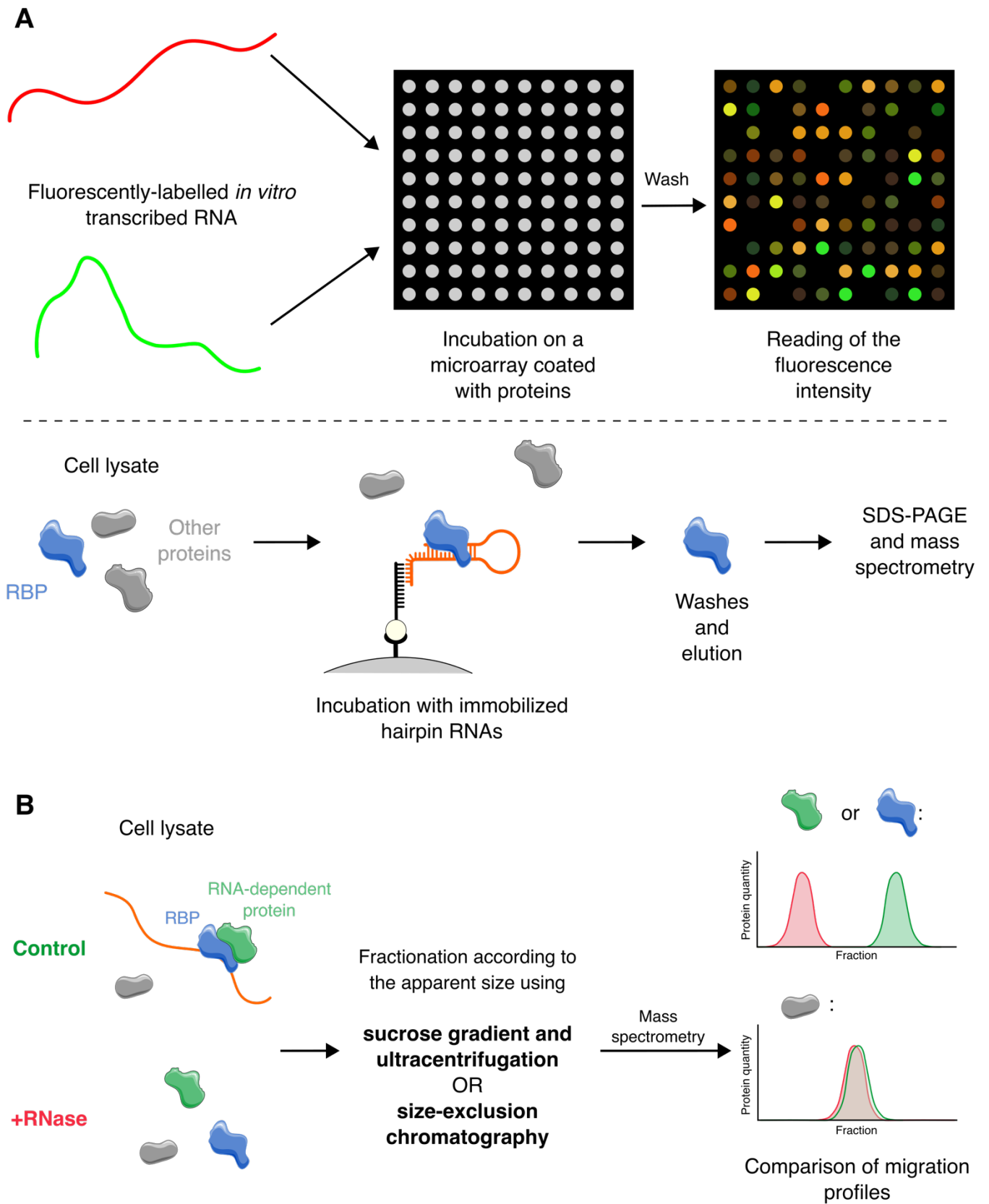


Figure 2: Schematic representation of the other RBP screening approaches
 A. Schematic representation of the RBP screens based on *in vitro* interactions. B. Schematic representation of the proteome-wide screens detecting RNA-dependent proteins.

In yeast, two studies used protein microarrays containing the majority of the proteome, and incubated them with fluorescently labeled total and *in vitro*-transcribed RNAs. The readout of the fluorescent signal allowed to quantify the interaction of each individual protein with the two different RNA samples (Figure 2A) (86, 87). Up to 80% of the yeast proteome could be analyzed at once with this technique (87). But the interaction between the RNAs and the protein only occurred *in vitro*, and could thus be biased by the absence of interacting partners or specific post-translational modifications (PTMs) that would be necessary for a protein to interact with its target RNA.

The same type of *in vitro* interaction study was performed using immobilized miRNA hairpins to pull-down interacting proteins. 72 different RNAs were immobilized and used as baits to interact with cell lysates. The associated proteins were then isolated by SDS-PAGE (sodium dodecyl sulfate polyacrylamide gel electrophoresis) and analyzed by mass spectrometry (Figure 2A) (86). This experiment has the advantage to screen for RBPs interacting with a specific RNA, but still presents the bias caused by the *in vitro* interaction.

Finally, two other techniques were developed around the concept of RNA-dependent proteins. Two lysates were compared to each other: a native cell lysate, and an RNase-treated cell lysate. The protein complexes of these lysates were then separated based on their apparent size, either on a sucrose gradient (89), or using size-exclusion chromatography (90). The different fractions were finally analyzed by mass spectrometry, to establish a migration profile for each protein in each of the two conditions (native lysate and RNase-treated lysate, Figure 2B). If the protein was part of an RNA-dependent complex, this protein should run at a higher apparent size in the native condition than in the RNase-treated condition, in which the complex had been disrupted and the protein was free. These methods made the detection of any protein that depends on RNA for its interactions possible, even if this protein does not bind to RNA directly. These proteins were accordingly called RNA-dependent proteins (89, 90). These techniques do not present any selection bias, and are quantitative. They also conserve the native interactions occurring in the cell, and hence can be useful to compare different cellular conditions. But they require more starting material than the methods presented before, which might not be easy to achieve for every biological sample.

1.2.4. Computational prediction of RBPs

Along with the development of experimental techniques to uncover new RBPs, several algorithms have been developed to predict RBPs based on different protein features. These prediction tools present the advantage to be inexpensive and rapidly applicable to different organisms compared to experimental approaches (91). Until 2019, 20 predictors restricted to RBP prediction were published, and 8 more could predict RNA-binding properties along other features such as DNA- and protein-binding (92).

The first approach, adopted by a majority of studies, is the prediction of RBPs or RNA-binding domains based on the sequence of the protein. Three components are often used to train a machine-learning model: the evolutionary conservation, the solvent accessibility and the propensity of amino acids for RNA-binding. More specifically, positively-charged residues are more likely to interact with the negatively-charged RNA backbone (93). An index of amino acid properties such as AAindex (amino acid index) can also be used (94). Thanks to the BLAST (basic local alignment search tool) and PSI-BLAST (position-specific iterative BLAST) programs, RNA-binding regions in proteins can be identified via the sequence similarity as well (95, 96). Finally, the evolutionary information is also interesting to predict RBPs, and the use of evolution conservation scores improves the performance of RBP predictions (91).

The majority of recent studies are now using both sequence and structural information to build prediction tools. The available structures from the PDB (protein databank) database allow to take the secondary structure of the protein into account, especially folding similarities between known RBPs and other proteins (91, 97). The 3D structure of the protein also permits the algorithm to account for the accessible surface area, which is especially interesting since RNA-interacting protein sequences are generally exposed on the surface of the folded protein. The number of resolved protein structures, especially of RBPs in complex with their target RNA, is however rather limited, but new data coming from powerful deep-learning algorithms such as AlphaFold could help fill this gap in knowledge (98, 99).

The main hurdle in developing structure-based machine-learning algorithms to predict RBPs is to define a positive and a negative dataset (100). Finding a reliable set of RBPs to establish a positive dataset for the algorithm to learn from can be hard, especially since the RNA-binding

regions are not known for many of newly discovered RBPs (100). Taking a very restricted dataset bears the risk of limiting the discovery of RBPs to proteins containing only the best known RBDs. On the contrary, taking a dataset that is too large will increase the rate of false positives. Furthermore, establishing a negative dataset can also be challenging, since it requires to have a subset of proteins known to never bind RNA but presenting enough sequence diversity. For all of these reasons, most algorithms that were developed in the last years focus on the binding to a specific RNA type, or try to incorporate more data such as the RNA secondary structure, to improve the specificity of the prediction (100).

The last type of data that can be used for the prediction of RBPs is relative to the interactions of the protein. One method that can be used is docking, a computational approach that relies on the component coordinates to evaluate the probability of an interaction between two proteins (101). No RNA-specific algorithm has been published to date, but several protein-protein ones have been adapted to be able to take RNA coordinates for one of the interactors. But this method only gives information for one protein-RNA couple at once (91). Another approach is to consider protein-protein interactions (PPI) that are already available, for example in the String database (102). A study of already annotated RBPs showed that these proteins tend to interact with each other more than with other types of proteins. Therefore, the more RBPs can be found in the PPI network of a given protein, the more likely is this protein to be an RBP itself. This principle has been applied in the SONAR algorithm (103).

1.2.5. New RBPs, new questions

All of the published experimental RBP screens yielded an important amount of data on RBPs in the last decade (24). Importantly, they detected already known “core” RBPs, involved in RNA metabolism, but also new RBPs, sometimes termed enigmRBPs. These proteins are able to bind RNA in addition to their already established function (24, 104).

This is notably the case for several metabolic enzymes, that were shown to “moonlight” as RBPs (24). One example is the IRP1 protein involved in iron metabolism. In cells loaded with iron, this protein binds to Fe-S clusters and has an aconitase activity. In the absence of iron, IRP1 binds to Iron Response Elements (IREs) in the 5' UTR or the 3' UTR of an mRNA, and inhibits the translation or the degradation of the mRNA, respectively. This mechanism allows for a quick response of the cell to a change in iron availability by acting at the post-

transcriptional level directly (105). Other proteins, such as GAPDH (glyceraldehyde-3-phosphate dehydrogenase), have been found to bind mRNAs and be able to regulate their translation (24).

This regulation of translation by non-canonical RBPs can also support the coordination between different cellular processes. Cyclin A2, a protein expressed during the S phase of the cell cycle and active in complexes with CDK1 and CDK2 (cyclin-dependent kinase 1 and 2), is necessary for the progression of the cells into mitosis (106). But this protein also binds the MRE11 (meiotic recombination 11 homolog 1) mRNA, and promotes its translation. In the absence of Cyclin A2, stalled replication forks caused by the lack of the MRE11 protein cause chromosomal instability. Thus, Cyclin A2 is able to coordinate the expression of proteins necessary for fork resolution and mitotic progression via the regulation of MRE11 mRNA translation (107).

However, this accumulation of data leads to the detection of more and more proteins as RBPs. In 2018, 1393 RBPs in 6 different datasets were reviewed (24). In 2020, the RBP2GO database compiled 43 RBP datasets in *Homo sapiens* alone, with 6100 proteins detected at least once as binding to RNA (108), which represents more than 4 times more RBPs within only two years. Furthermore, few overlaps can be observed between these datasets, with most of the proteins detected as RBP candidates in only one study (Figure 3, red bars). This can also be observed for *Mus musculus*, *Saccharomyces cerevisiae* and *Drosophila melanogaster*, and raises the question of the specificity of these proteome-wide RBP screens. Finally, many of the new RBPs detected also lack a defined RNA-binding domain, underlining the necessity for more research on this topic (24, 68, 75).

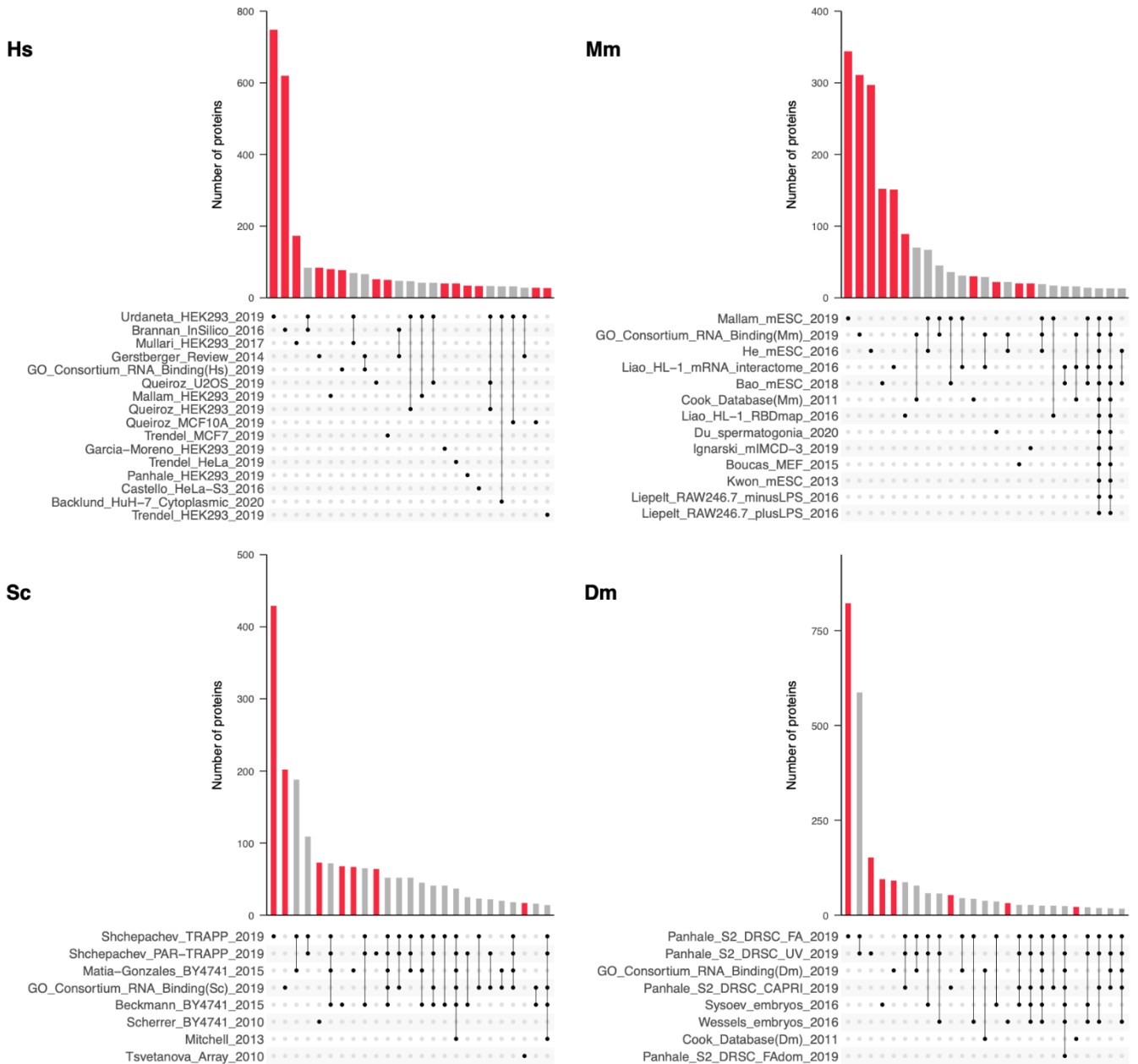


Figure 3: Upset plots showing the number of proteins detected as RBPs in one or more datasets in *Homo sapiens* (Hs), *Mus musculus* (Mm), *Saccharomyces cerevisiae* (Sc) and *Drosophila melanogaster* (Dm)

Only the first 25 intersections are shown. The red bars show the proteins present in only one dataset.

1.3. How do RBPs bind RNA: the RNA-binding domains

1.3.1. The classical RBDs

1.3.1.1. The different types of classical RBDs

Until the 2010s, it was thought that proteins bind to RNA via a small number of well-known canonical RBDs, whose sequence and structure has been known since the 1990s.

The most abundant of these RBDs is the RNA-recognition motif, or RRM. It is present in 0.5 to 1% of all human proteins, and is the most studied RBD to date (109). It can bind ssRNA (single stranded RNA) through its $\beta\alpha\beta\beta\alpha\beta$ topology, the RNA being generally bound on the surface of the β sheets. Two consensus sequences, called RNP1 and RNP2, are present in these sheets and recognize between 4 and 8 nucleotides on the target RNA (110, 111). Consequently, the recognition specificity of one isolated RRM is low (111).

The K-homology domain, or KH domain, was named after the hnRNPK protein, where it was first discovered. This 70 amino acids (aa) long domain also possesses three α helices and three β sheets, their topology making the difference between subtypes 1 and 2. It recognizes ssRNA, and sometimes ssDNA (112). More specifically, a consensus domain, the GxxG loop, can recognize 4 nucleotides on the target RNA, conferring to this domain a low specificity when considered alone (111, 112).

The double-stranded RNA-binding domain (dsRBD) is 70-90 aa long, with an $\alpha\beta\beta\beta\alpha$ structure. Contrary to the RRM and KH domains, it recognizes only structured double stranded RNAs (dsRNAs), in majority via contact with the 2'-OH groups and phosphate backbone of the RNA molecule. However, the exact binding mode is not the same for all proteins containing a dsRBD. The specificity of the binding is ensured by several dsRBDs that collaborate to recognize specific 3D shapes. The dsRBD have also been found to bind to proteins and dsDNA (109, 113).

Other domains, such as the S1 domain, are well known. This domain, which recognizes RNA via two β sheets in a similar fashion as the RRM, was discovered in the ribosomal protein S1. Since then, it was found in other RBPs. Other well-known domains, such as the PAZ and PIWI domains, are present in specific RBPs and can bind very specific RNAs. These domains are found in the proteins involved in the processing of piRNAs and miRNAs (109).

Finally, some domains known to bind other type of ligands can also bind RNA. Zinc fingers are domains containing a zinc ion and originally discovered as DNA-binding domains. There are several subtypes, differentiated by the residues coordinating the zinc, that have also been shown to bind RNA, with different modes of recognition implicating different arrangements of the fingers on the RNA molecule (109). Different fingers can either recognize some bases or bind to the backbone of the RNA, as it is the case for the TFIIIA (transcription factor IIIA) (114).

1.3.1.2. Classical RBDs are often arranged in a modular fashion

All of the globular domains mentioned above are able to bind RNA individually, but possess a very low specificity, often recognizing less than 10 nucleotides on their RNA target. But in most of the RBPs, several copies of these domains are present. This increases the binding surface on the protein, and hence the sequence recognized by this surface, and in turn improves the specificity of the RBP for a certain subset of RNA targets. The RBP can also recognize a specific RNA 3D conformation with several RBDs (109, 111).

The same RBD can be present in several copies in the same protein. For example, the TFIIIA transcription factor possesses nine CCHH zinc fingers in a row, some of them binding to DNA while some others bind to RNA (114). Similarly, the Staufen protein contains five dsRBDs, several of them being indispensable to recognize the shape of its dsRNA target (115). Several different RBDs can also collaborate in the same RBP. The Dicer human protein is involved in the microRNA (miRNA) pathway; it binds and processes miRNA precursors to produce mature miRNAs (116). The protein contains a PAZ domain as well as a dsRBD, separated by the endonuclease domains. While the PAZ and endonuclease domains are sufficient to bind and process miRNA precursors (117), the absence of the dsRBD domain reduces *in vivo* processing rates (118). Thus, the PAZ domain is responsible for the binding and positioning of the miRNA precursor, but the dsRBD enhances the RNA-binding capacities of the Dicer protein (118). The recognition of RNA by RBDs can also involve other protein domains. The SF1 (splicing factor 1) protein binds to the intron branch point sequence (BPS) during the assembly of the spliceosome. While the 3' part of the BPS is recognized by the KH domain of SF1, the 5' part is bound by its QUA2 (Quaking homology 2) region (119).

This modularity of RBDs allows a variety of different RBD combinations with simple building blocks, increasing the specificity of RBPs and the amount and complexity of their potential targets. The establishment of several weak interactions on an RNA target also facilitates the assembly and disassembly of RNPs, and supports the fine tuning of the RNA-binding capacities of a specific RBP, since each RBD can be regulated separately (109).

1.3.2. The expansion of RBPomes unveiled new RBDs

Less than a third of the RBPs identified in the last decade by proteome-wide screens contains one of the canonical RBDs that were just described (120). In addition, studies aiming at detecting RNA-binding regions within RBPs uncovered new non-canonical RBDs, notably domains known for other functions that can also bind to RNA (24, 75).

Heat shock proteins (HSPs) were discovered because of their overexpression during the heat shock response. They are molecular chaperones, helping other proteins to fold into an active conformation, or to interact with their ligands and other proteic partners. HSP90 is one of the major molecular chaperones, and has a very important role in the response to multiple stresses (121). Its N-terminal domain forms a clamp that binds to the protein's ligands (122). This domain was also detected as a new RBD in several studies, notably in human and in *Drosophila* (75, 123). Furthermore, other studies showed that HSP90 mediates the loading of RNA duplexes into the Argonaute protein (124). HSP70, another heat shock protein, has also been detected as binding RNA (75). This protein was shown to bind AU-rich elements in mRNAs with its ATPase and peptide-binding domains, stabilizing the bound transcripts (125). This mRNA-stabilizing activity is independent from its chaperone activity (126). These results indicate that molecular chaperones and their domains involved in ligand-binding can also bind to RNA as a separate function. It underlines the fact that other domains than the canonical RBDs can be involved in the binding and regulation of RNAs.

1.3.3. The role of disorder in RNA-binding

Another type of domain that was detected as RNA-binding in several studies is intrinsically disordered regions (IDRs) (75, 123). These protein regions are defined as such because they do not adopt any specific conformation in the folded protein. They are generally enriched in positively charged amino acids such as lysine and asparagine, as well as negatively charged amino acids like aspartic acid and glutamic acid, and in tyrosine (24, 127, 128).

As was shown in several proteome-wide RBP screens and bioinformatic analyses, RBPs are enriched in IDRs compared to the rest of the proteome (68, 75, 120). This enrichment is higher in proteins containing canonical RBDs compared to proteins containing non-canonical RBDs (120). Disordered regions can indeed be found in between canonical RBDs, generally in the form of long and flexible linkers. (109). For example, the ADAR2 (adenosine deaminase RNA specific 2) protein binds dsRNA via two dsRBDs separated by a 90 nucleotide-long disordered sequence. The IDR facilitates the interaction of the protein with RNA by conferring an additional flexibility to the protein (129, 130). IDRs thus play a role in RNA-binding.

Specific IDRs can also bind RNA directly. For example, basic arms are clusters of basic patches often observed in RBPs. They are composed of 4-8 lysines or arginines forming a highly positive and exposed interface. They form "basic islands", flanking canonical RBDs, or alternating with acidic patches in a repetitive manner (127). Arg-rich motifs, or ARMs, have been mostly described in viral proteins, like the Tat protein, and confer them unspecific RNA-binding features (131). Poly-lysine or poly-K peptides are able to bind RNA *in vitro*, and poly-K patches have been identified in RBPs lacking canonical RBDs such as SDAD1 (SDA1 domain-containing protein 1) (132).

RG-rich repeats, also called RGG-box or GAR repeats, have been known since the 90s as RNA-binding motifs in the hnRNP proteins, and have been recently classified as IDRs (133). They can be divided in three categories: di-RG, tri-RG and RGG repeats, and each repeat type can be present from tens of copies to thousands of copies (134). In hnRNPU, the RGG-box binds to the Xist lncRNA (135). The FMRP protein is also able to bind RNA via RGG-repeats; the RGG patch binds to a G-quadruplex as well as surrounding sequences on the sc1 RNA (136). The last type of IDR that can bind to RNA is the RS repeat. RS repeats are repeats contained in SR or SR-like proteins, often in combination with one or more RRM. These proteins are involved in splicing enhancement and other steps of RNA metabolism. RS repeats are able to bind RNA directly and promote adjacent intron splicing (127). The SRSF1 (serine and arginine-rich splicing factor 1) protein possesses two RRMs and an RS domain comprised of 8 RS repeats, which enhances the affinity of the RRMs for their target RNA. The RS domain can be phosphorylated and will then transition into a folded conformation, which increases the RNA-binding ability of the protein (137).

As was already mentioned, IDRs can be present in proteins with canonical RBDs. The FUS protein contains a zinc finger domain, an RRM and RGG-repeats in its prion-like domain. While the RRM and the zinc finger can bind RNA on their own, RGG repeats also bind to the RNA target. This increases the binding affinity and promotes the unfolding of the RNA structure, uncovering additional binding sites in the target RNA (138).

Nevertheless, some RBPs only contain IDRs and no other RBD. In turn, RBPs are enriched in proteins that are highly disordered (more than 80% of their sequence) (120). The NF-kappa-B-activating protein, or NKAP, is involved in transcriptional repression and Notch-mediated T-cell development. Almost 75% of its sequence is disordered, comprising an RS motif and a basic patch, and the only other domain it contains is a DUF 926 domain. It was recently shown to have a role in splicing and to be present in nuclear speckles. NKAP interacts with the U1, U4 and U5 small nuclear RNAs, and its IDR is required for its localization in nuclear speckles as well as its RNA-binding ability (139). Furthermore, highly disordered proteins are more susceptible to undergo LLPS, and aggregate in membrane-less organelles (140). As was previously mentioned, this process can be regulated by RNAs binding to the phase-separating proteins, showing a role of IDRs in the regulation of some protein behavior by RNA (34).

Disordered regions are difficult to study using standard structural biology methods because of their dynamic folding behavior, and few resolved structures are available. Prediction algorithms have been developed in the last two decades to circumvent this problem, the first one being DisEMBL (141). Most of these algorithms give a propensity for disorder as an output, and the user can then decide which threshold they want to use to define IDRs. More recent algorithm, such as MobiDB-Lite, combine several different algorithms with experimental data from PDB to make a prediction on disordered regions (97, 142). The results of MobiDB-lite as well as other algorithms and resources can be found on the MobiDB database (143). Some widely used databases such as Uniprot and InterPro recently started to provide access to such predictions, facilitating the access to this data (144, 145). However, the evaluation and comparison of the different algorithms in the CASP10 (critical assessment of methods for protein structure prediction, 10th round) and more recently in the CAID (critical assessment of intrinsic protein disorder prediction) showed that while some algorithms lacked specificity, others were too conservative and missed some well-known disordered proteins (146, 147). More recently, AlphaFold showed that the regions for which no reliable structure can be predicted overlap with known and predicted disordered regions, adding another source

of information on IDRs (98). Some algorithms can furthermore predict functions for IDRs, such as RNA-binding, but no prediction of the different subtypes of disorder that were mentioned earlier could be found.

1.4. Aim of the study

The accumulation of data on RBPs in the recent years yielded incredible discoveries, but the overlap between the different datasets that were published is small, in human as well as mouse or drosophila. Additionally, more and more proteins are detected as RBPs but lacking any defined RBD, raising the question of which RNA-binding features they may contain. Are those proteins mainly containing unknown RBDs, or disordered regions? Can the presence of an RBD be used to better distinguish relevant RBP candidates and make the navigation in this wide pool of data easier for scientists studying RBPs?

Hence, the aim of my project is to study the presence of RNA-binding features in the RBPs compiled in the RBP2GO database. It can be divided into the following objectives:

- to compile a list of known RBDs, as well as other RNA-binding features and study the repartition of these features in the RBPs across the 13 different species present in the RBP2GO database
- to establish the importance of the listed RNA-binding features for the characterization of RBP candidates
- to upload the acquired data in the RBP2GO database and provide an easy access to it.

To address these questions, I took advantage of the data compiled in the RBP2GO database, and mined the literature as well as the InterPro database for known RNA-related annotations (108, 145). After a selection of RBDs and RNA-related family InterPro IDs (Rfam IDs), I looked at the repartition of these annotations in the proteins of the database. I also extracted data from the MobiDB database to study the presence of disorder in these proteins and the correlation with experimental findings (143). New RBP and RBD candidates were also investigated based on the acquired knowledge on RBDs and Rfam IDs, and a new score for the evaluation of RBP candidates was established. Finally, all the information on RNA-related annotations and disorder gathered in this analysis, as well as the new RBP2GO composite score, were added to the RBP2GO database, to facilitate the selection of RBP candidates by the users (Figure 4).

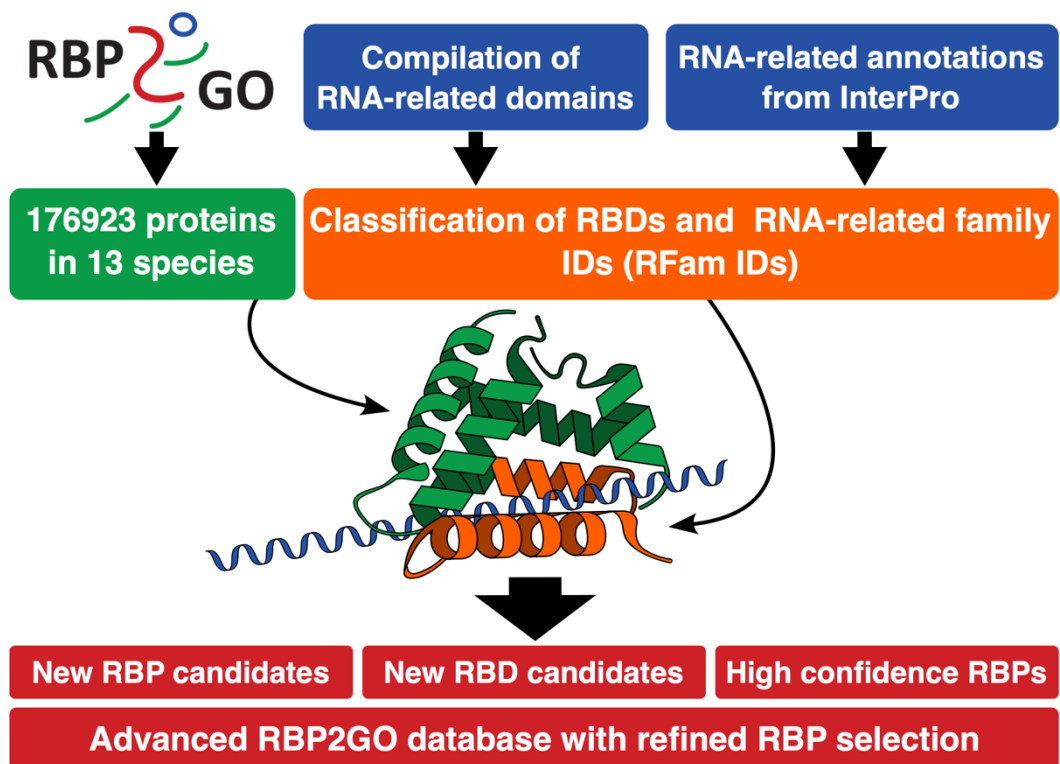


Figure 4: Graphical abstract summarizing the study presented in this thesis

2. Results

2.1. Selection of the RNA-binding domains from the literature and the InterPro database

To study the presence of RBDs in the proteins of the RBP2GO database, a list of known RBDs was necessary. I compiled three published lists of RBDs: the list of canonical RBDs from Gerstberger et al. (148), as well as the two lists of newly discovered RBDs from Castello et al. (68, 75), for a total of 809 IDs. To ensure I had a comprehensive list, I also searched for the terms “RNA-binding” in the InterPro database (145) and selected all InterPro IDs matching this search. InterPro also contains an “RNA-binding domain superfamily” (IPR035979), so I added to the list all InterPro entries overlapping with this ID. Finally, the InterPro IDs are also manually labelled with GO (Gene Ontology) terms (149), so all IDs annotated with the “RNA binding” GO term (GO:0003723) were selected as well. This yielded 2252 additional InterPro IDs. All IDs were compiled together, and only the “Domain” and “Repeat” types of IDs were selected (Figure 5A).

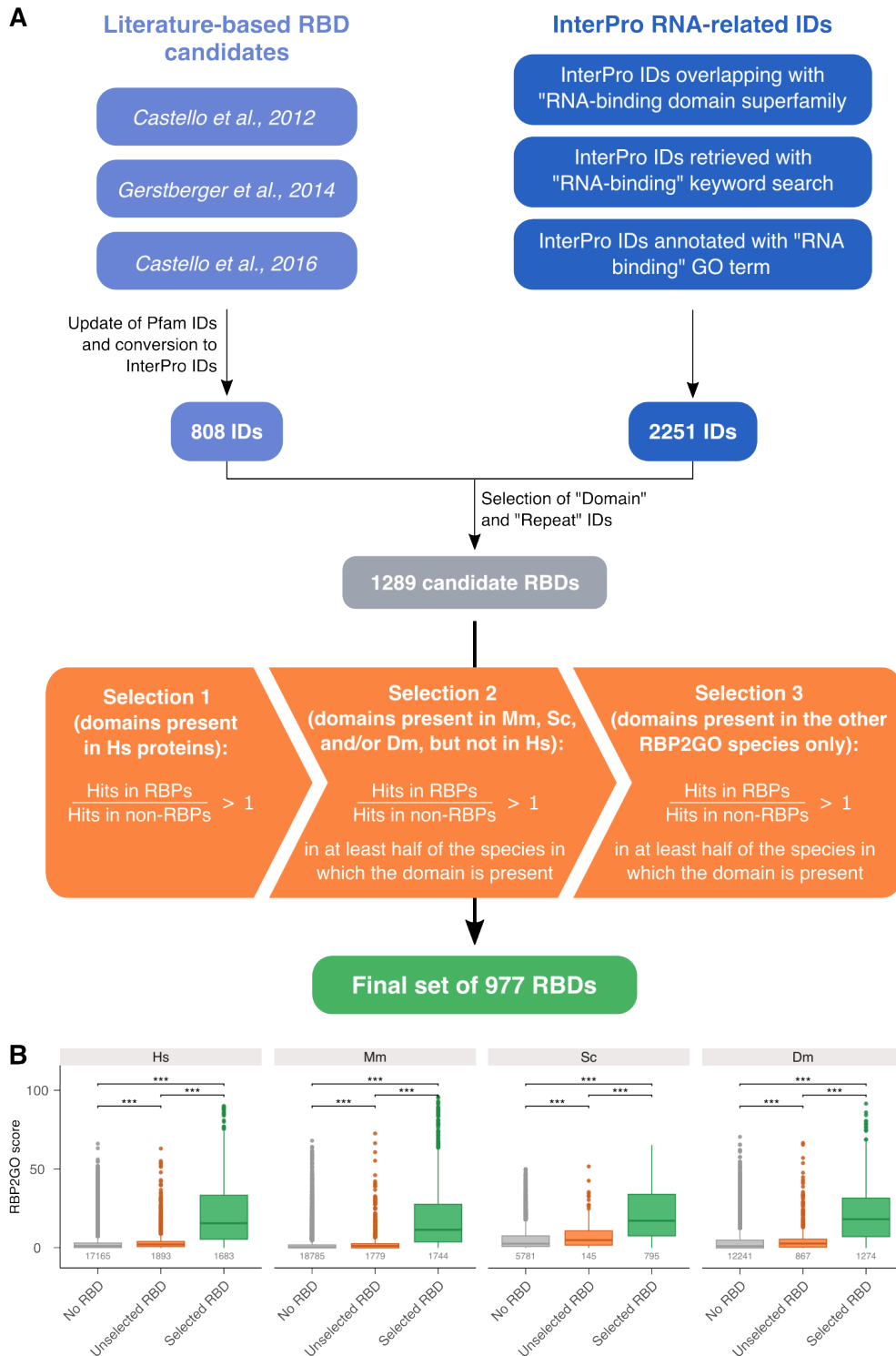


Figure 5: Selection process of the RNA-binding domains.

A. Flow chart showing the selection process of the RBD candidates and depicting the three-step selection procedure based on the ratios between RBPs and non-RBPs in the species available in the RBP2GO database (20). The starting lists of RBD candidates and the final list of selected RBDs are found in Supplementary Table S1 and Supplementary Table S3, respectively. B. Boxplots depicting the distribution of the RBP2GO score of the proteins from the groups with no RBDs (grey), unselected RBDs (orange) and selected RBDs (green) in Hs, Mm, Sc and Dm. The numbers below the boxplots indicate the size of the groups. *, **, *** and **** correspond to p-values < 0.05, 0.01, 0.001 and 0.0001, respectively, resulting from a Wilcoxon rank sum test.

RBP2GO also provides the number of datasets in which a given protein has been detected as RNA-binding (108). This data serves to calculate an RBP2GO score for nine species out of 13 for which sufficient data is available: the first half of the score corresponds to the proportion of datasets detecting this protein as RBP, normalized to 50, and the second is the mean of this number for the top 10 STRING interactors of the protein (102). This results in a score reflecting the likelihood of an RBP candidate to indeed bind RNA *in vivo*, taking into account that RBPs preferentially interact with other RBPs (103). I used this information to calculate the median RBP2GO score for the proteins with no RBD, the proteins with unselected RBDs and the proteins with selected RBDs. The proteins with the selected RBDs exhibited a significantly higher score than the two other groups in the four most studied species. The proteins with unselected RBDs also had a significantly higher score than the proteins with no RBD, but the difference was visibly smaller (Figure 5B). These results validated the selection process, and the selected InterPro IDs are hereby further referred to as RBDs.

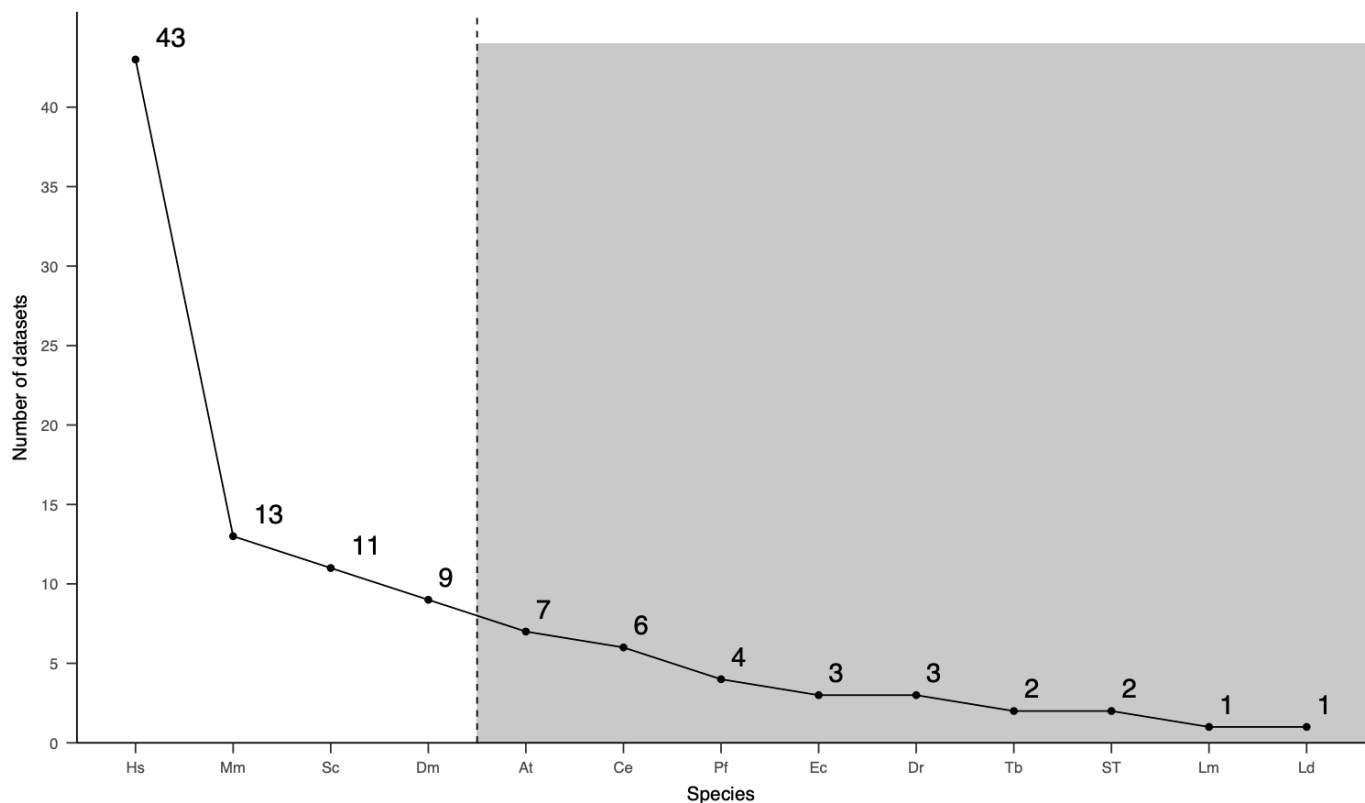


Figure 6: Scatterplot displaying the number of datasets compiled in the RBP2GO database for each species

The number above each point displays the number of datasets compiled in RBP2GO for the given species. The species in the white area, before the dashed line, are referred to as “well-studied” species.

2.2. Distribution of the selected RBDs in the species and the proteins of the RBP2GO database

Once the list of RBDs was established, I analyzed the repartition of these InterPro IDs in the proteins of the different species compiled in the RBP2GO database. The RBDs were present either in RBPs only, non-RBPs only or both types of proteins for each species. The well-studied species (above the dashed line) also had the lower proportion of RBDs present only in non-RBPs. Of the less studied species, only Ec and Pf showed a similarly low proportion of RBDs present in non-RBPs only. The proportion of RBDs present only in RBPs was lower for the less studied species; for example, while more than 65% of the RBDs present in Mm can be found in RBPs only, this is the case for less than 7% of the RBDs present in Dr (Figure 7A).

When comparing the proportion of RBPs and of RBD-containing proteins in the proteome in each species, again a discrepancy between highly and lowly studied species, except for Pf and Ec and Lm, can be found. While almost 30% of all human proteins were found at least once to be binding to RNA, less than 10% of the proteome actually contains an RBD. For At, Ce and ST, the proportions of the proteome constituted by RBPs and RBD-containing proteins looked similar, while for Dr, Tb and Ld, there were far less RBPs than RBD-containing proteins (Figure 7B). The same phenomenon was observed when looking, for each of the top six most abundant RBDs, at the proportion of RBPs in the proteins containing these domains. All of these domains are classical RBDs, with the RNA-recognition motif being the most abundant of all RBDs. Not all of these domains are present in all species, such as the RRM which is absent in Ec and ST. But nevertheless, for the other five species with less than 4 datasets compiled in RBP2GO (Figure 6), less than 50% of the proteins containing these canonical domains were listed as RBPs in RBP2GO (Figure 7C).

Taken together, these results showed a discrepancy in the repartition of the RBDs between the proteins of the well-studied and lowly-studied species. While the species with the most studies exhibited most of their RBDs in RBPs and more RBPs detected than RBD-containing proteins, the reverse situation was observed for lowly-studied species, with the notable exception of Ec. Thus, a low number of proteome-wide RBP datasets resulted in a lack of coverage, and some RBPs were likely missed in species covered in fewer studies.

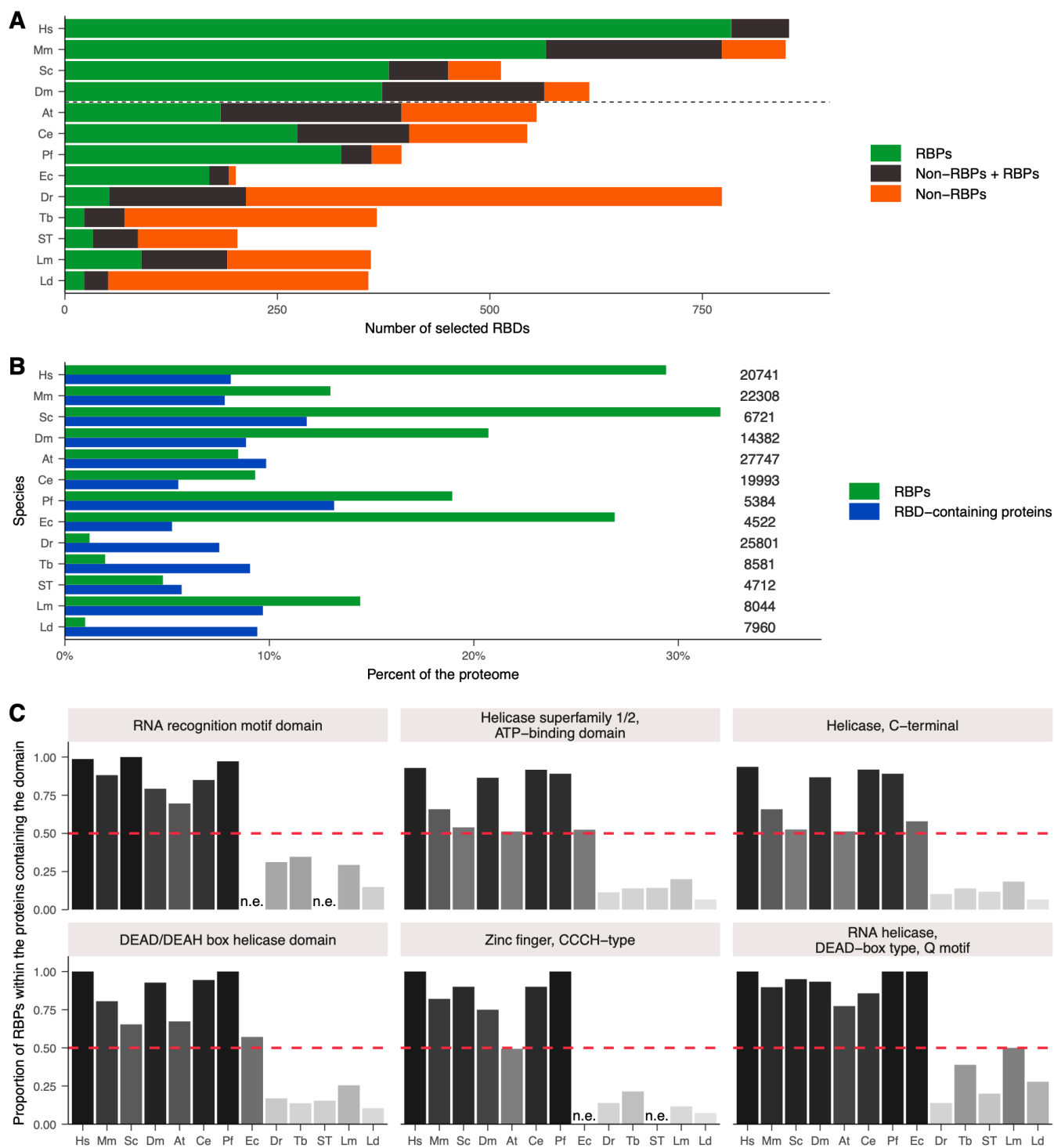


Figure 7: Analysis of the distribution of RBDs in the species of the RBP2GO database

A. Number of selected RBDs per species grouped according to their distribution in RBDs present only in RBPs (green), RBDs present in both non-RBPs + RBPs (black) and RBDs present only in non-RBPs (orange). B. Proportion of the proteome represented by RBPs (green) and RBD-containing proteins (blue). The numbers at the end of the bars indicate the size of the respective proteomes. C. Proportion of RBPs per species in the proteins containing the indicated RBD. The high percentages are represented in dark grey, and the low percentages in light grey. The domains are ranked from more abundant in the RBP2GO proteins to the least abundant. n.e. = non-existent, i.e. the given domain is absent from this species.

To further explore this situation, I studied the proportions of RBPs and non-RBPs, containing or not an RBD, in the proteomes of the 13 species. Surprisingly, the most studied species such as human harbored a higher percentage of RBPs with no RBD compared to RBPs with RBD. RBPs with RBDs represented 7% of the human proteome, while RBPs with no RBD made up 22%. This is also the case for Mm, in which only a third of the RBPs harbor an RBD (Figure 8A). Less surprisingly, the least studied species had a higher proportion of non-RBPs with RBDs. It ranged from 7.8% to 9.2% of the proteome for Dr, Tb, ST, Lm and Ld, compared to 1.1% to 2.9% for Hs, Mm, Sc and Dm. This confirmed that some RBPs were certainly not detected in these species, due a lack of proteome-wide RBP studies (Figure 8A). Vice versa, some non-RBPs containing RBDs could be found in all species, notably 231 in human, and raised the question of why they were not detected in any of the 43 datasets available for Hs on RBP2GO. Data from a mass spectrometry analysis in HeLa cells was downloaded from the EBI website to search for an answer (150). It appeared clearly that RBPs exhibited a significantly higher expression than non-RBPs, while the presence of an RBD did not significantly influence the expression of the proteins (Figure 8B). Furthermore, there was a positive correlation between the RBP2GO score and the expression of the proteins in HeLa cells (Figure 8C), implicating that the non-RBPs with RBD may not have been detected in proteome-wide screens due to their low expression.

Overall, these results showed a difference in the repartition of RBDs in the species of the database. The species with the highest number of datasets available displayed most of their RBDs in RBPs, but they also showed a large proportion of proteins detected as RBPs with no RBD. On the other hand, the species with less than four datasets available had a high proportion of RBDs in non-RBPs, suggesting some RBPs were not detected in the proteome-wide screens.

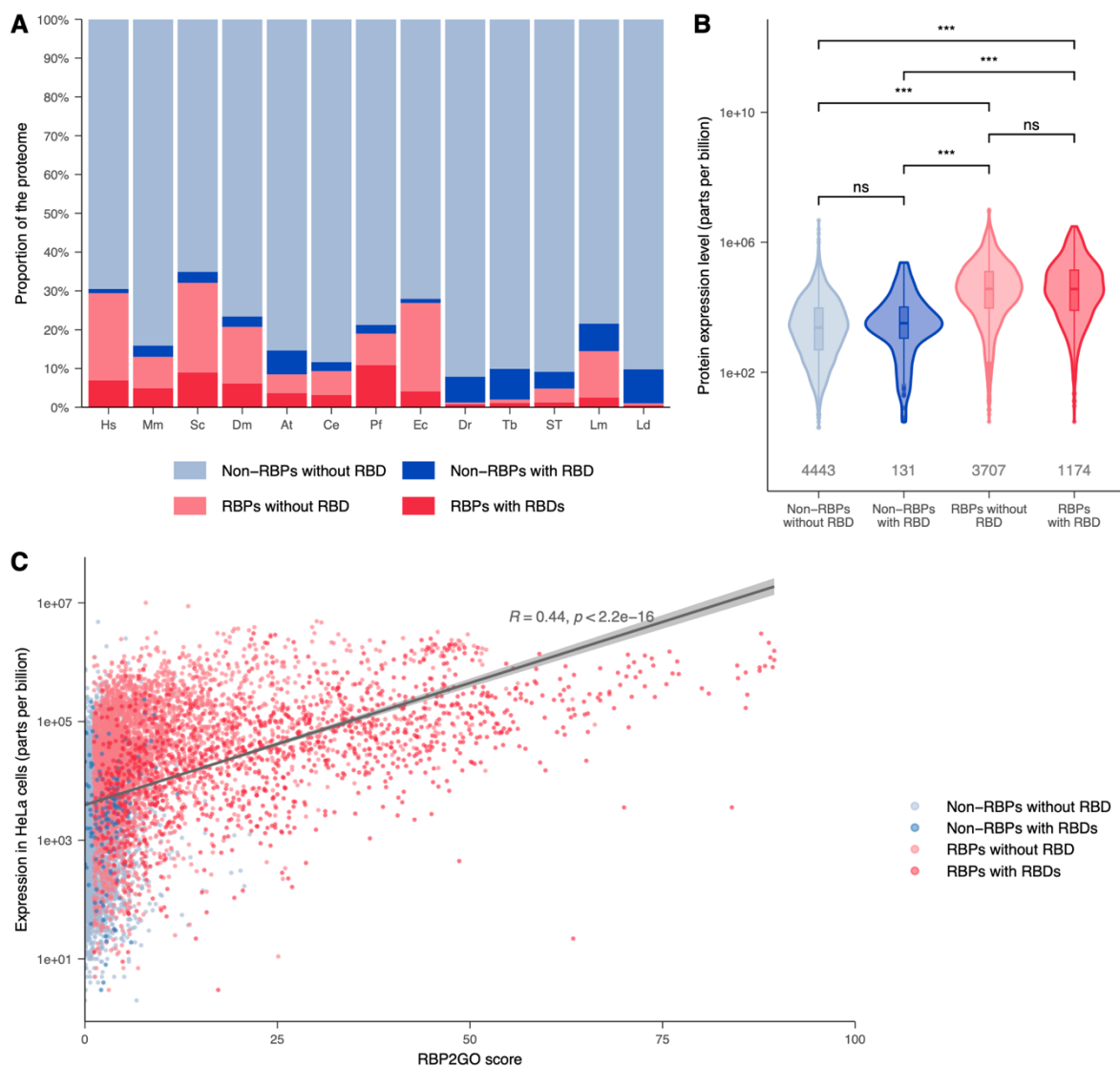


Figure 8: Distribution of the RBDs in the RBPs and non-RBPs of the RBP2GO database and correlation with their expression

A. Proportion of the proteome in % of non-RBPs without RBD (light blue), non-RBPs with RBD (dark blue), RBPs without RBD (light red) and RBPs with RBD (dark red) in the different species. B. Boxplots representing the protein expression level in HeLa cells according to a mass spectrometry experience (148) in the four protein groups: non-RBPs without RBD (light blue), non-RBPs with RBD (dark blue), RBPs without RBD (light red) and RBPs with RBD (dark red). The numbers in grey represent the number of proteins detected in the study for each group. *, **, *** and **** correspond to p-values < 0.05, 0.01, 0.001 and 0.0001, respectively, resulting from a Wilcoxon rank sum test. C. Scatterplot of the expression of human (Hs) proteins in HeLa cells and their RBP2GO score. R represents the Pearson's correlation factor.

2.3. Relevance of the presence of RBDs for the selection of RBP candidates for further validation

The presence of a known RBD in a protein has already been used to characterize RBPs (151), thus I explored the relationship between the presence of an RBD and the ability of the protein to bind to RNA. Since the RBP2GO database already provides the RBP2GO score to evaluate the RNA-binding potential of a protein, I used this summary of experimental knowledge to assess the impact of the presence of a selected RBD. When comparing the median score of proteins that contained or not an RBD, a clear and significant difference could be seen. The proteins harboring an RBD had a higher RBP2GO score than the proteins that did not contain any (Figure 9A). When dividing the proteins into four groups between non-RBPs and RBPs with or without an RBD, there was also a significant difference in all species; whether it is for RBPs or non-RBPs, the proteins with an RBD displayed a higher RBP2GO score (Figure 9B). Regarding the non-RBPs with an RBD, it further confirms that some RBPs were not detected in the proteome-wide screens.

This correlation between the presence of an RBD and the RBP2GO score was further explored by studying the proportion of RBD-containing proteins for each unit of the RBP2GO score (Figure 10). In most species, except *Ec*, *Dr* and *Tb*, there was a clear positive relationship between the percentage of proteins with an RBD and the RBP2GO score up to a score of 50 (black dashed line). After a score of 50, almost all of the proteins contained at least one RBD. For *Ec*, the plateau seemed to be reached at a score of 65. For *Dr*, the plateau was reached much earlier, at a score of 20, and no plateau could be seen for *Tb*, which also exhibited a much lower correlation coefficient. These observations strengthen the conclusion that the RBP2GO score, and thus the RNA-binding potential of an RBP, is positively correlated to the presence of a known RBD.

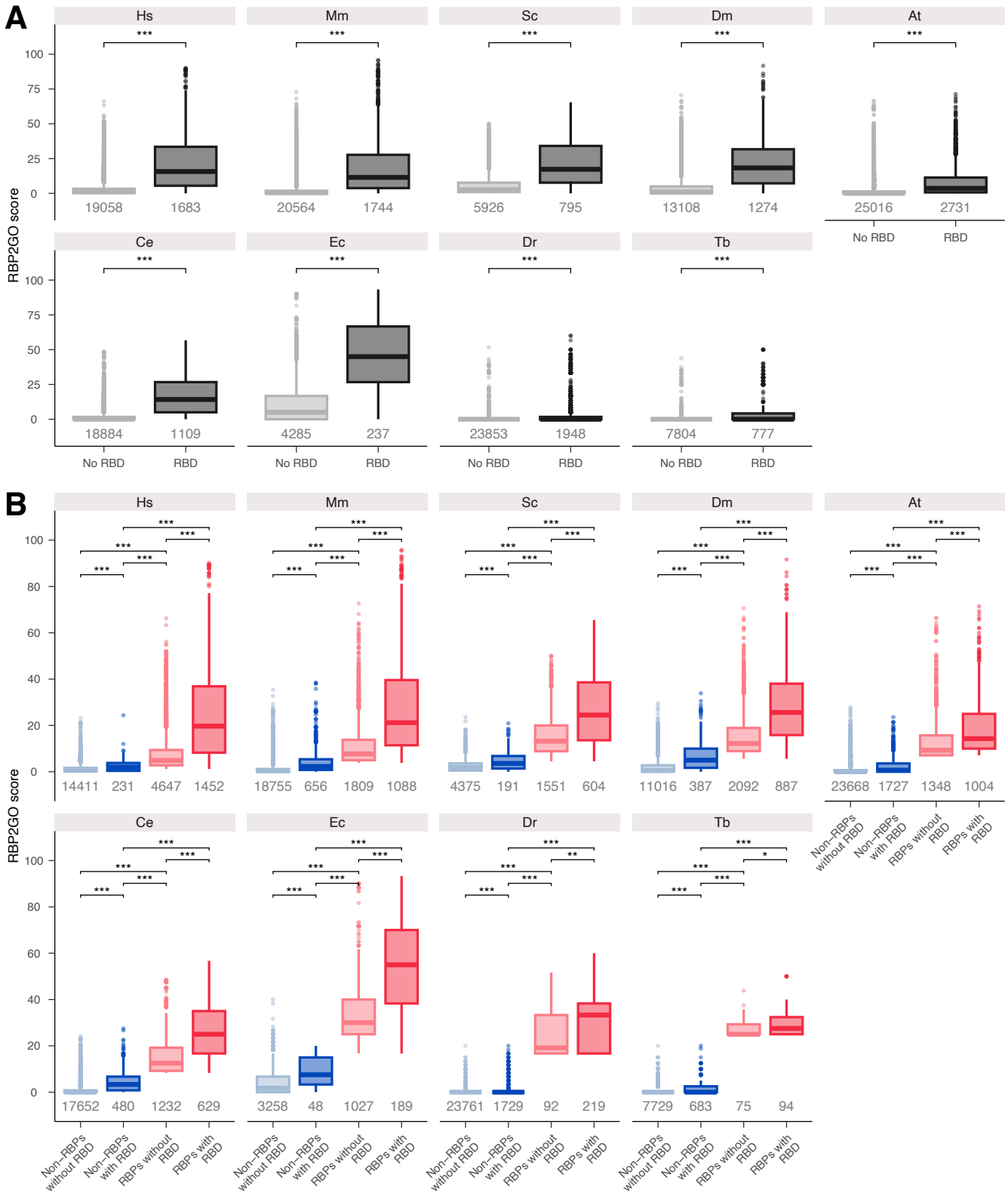


Figure 9: Influence of the presence of RBDs on the RBP2GO score

A. Boxplot representing the distribution of the RBP2GO score in proteins without RBD (no RBD, light grey) and with RBD (RBD, dark grey). B. Same as in A., but separated into non-RBPs without RBD (light blue), non-RBPs with RBD (dark blue), RBPs without RBD (light red) and RBPs with RBD (dark red). *, **, *** and **** correspond to p-values < 0.05, 0.01, 0.001 and 0.0001, respectively, resulting from a Wilcoxon rank sum test.

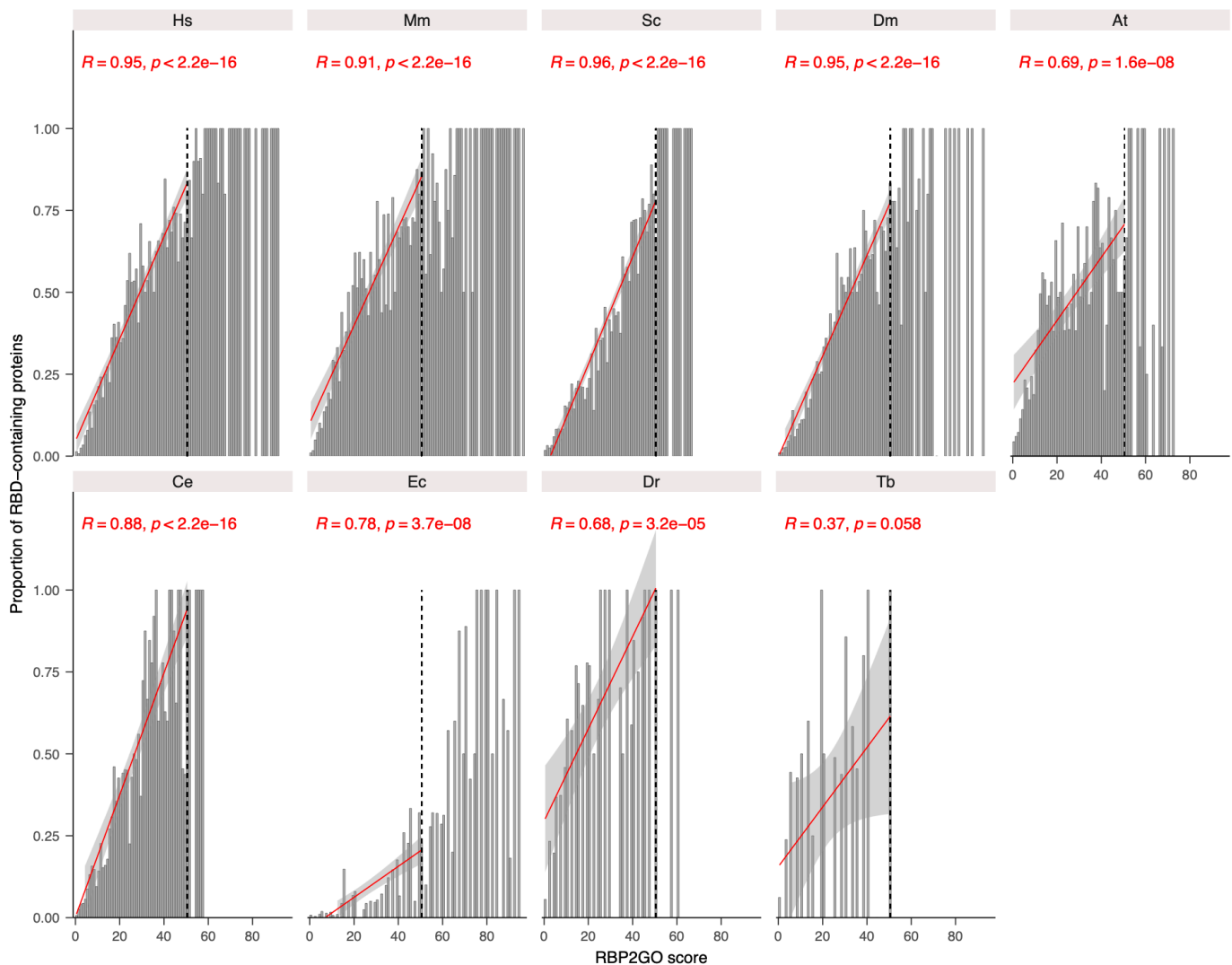
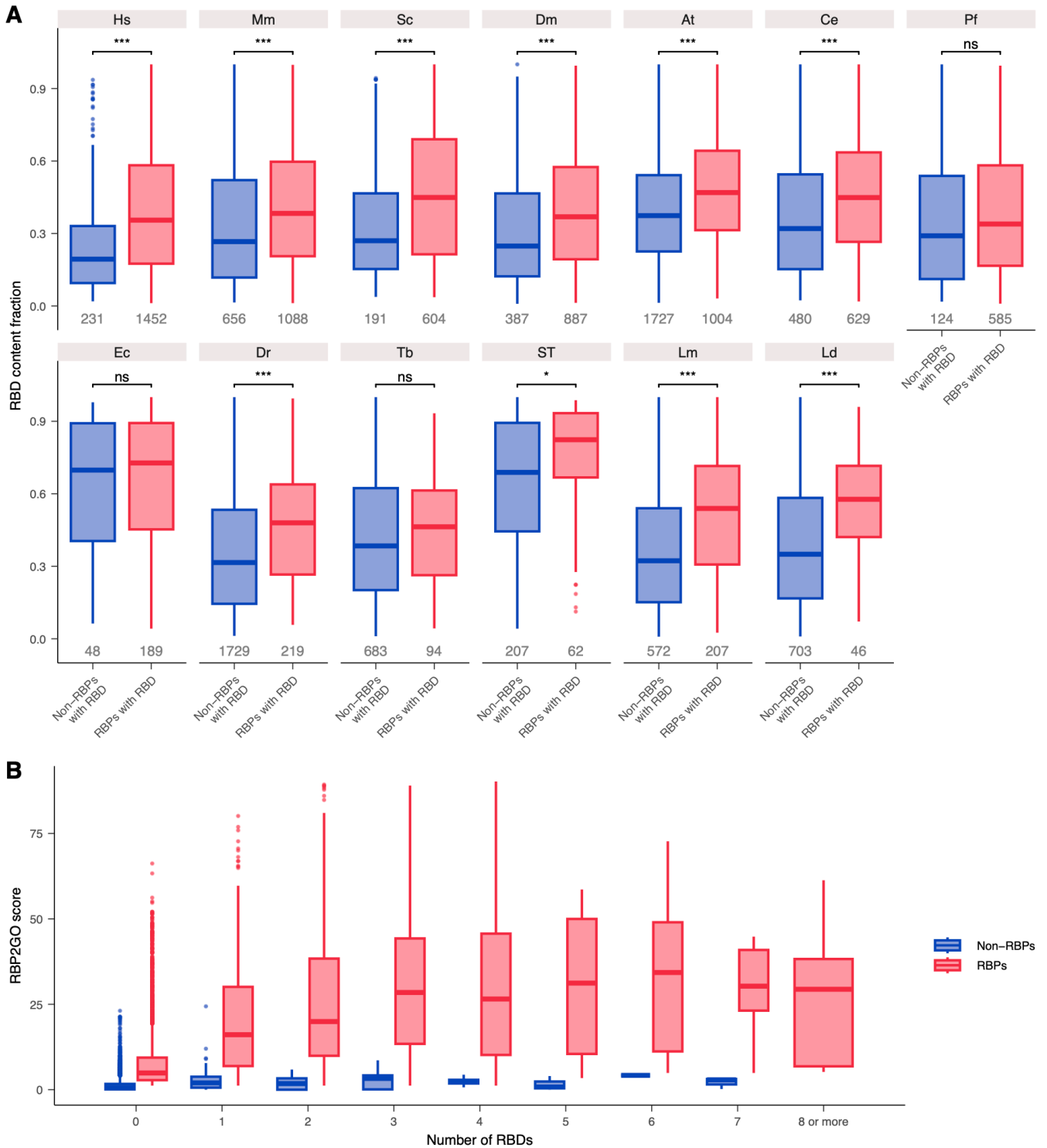


Figure 10: Proportion of RBD-containing proteins (in %) for each unit of RBP2GO score. The red line represents a linear regression between 0 and 50. R represents the Person's correlation coefficient.

The InterPro database also gives access to the coordinates of all InterPro annotations for every protein. This allowed me to calculate the percentage of the protein sequence, or content fraction, covered by RBDs, as well as the number of RBDs per protein. Interestingly, when analyzing the RBD content fraction of the proteins with RBDs, the RBPs showed a higher RBD content fraction than the non-RBPs (Figure 11A). This difference was significant in all species, except again for Pf and Ec as well as Tb. For the last two, the absence of significance could be attributed to the low number of non-RBPs or RBPs containing RBDs, respectively. In human, the median RBD content fraction in non-RBPs with an RBD was 19.4%, while it amounted to 35.5% in RBPs containing an RBD. Moreover, the more RBDs were present in a RBP, the higher its RBP2GO score was for human proteins, while this relationship was not observed for non-RBPs (Figure 11B).



Overall, the presence of an RBD, but also the RBD content fraction and the number of RBDs present in a protein correlated positively with its RBP2GO score. This points towards a correlation between the RNA-binding potential of a protein *in vivo* and the presence of RBDs, and thus a potential use of RBDs for the refinement or prediction of RBPs. However, more than two thirds of human RBPs lack an identified RNA-binding feature, and this is generally observed for all well-studied species.

2.4. Selection of the RNA-related annotations from the literature and the InterPro database

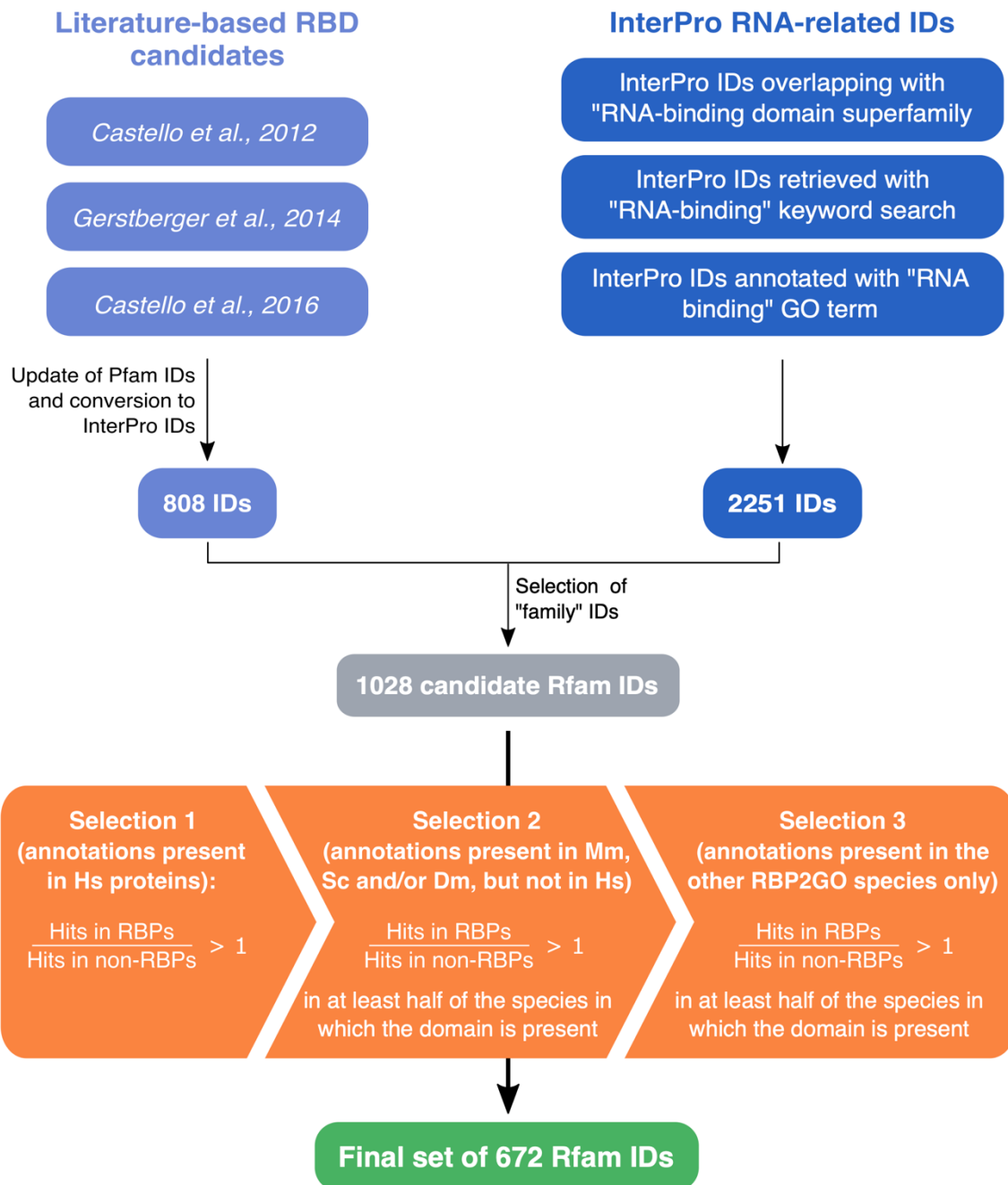


Figure 12: Selection process of the RNA-related family IDs (Rfam IDs)

In addition to the presence of domain annotations, the InterPro database also provides information on protein families in the form of family IDs. These annotations do not describe particular domains with coordinates, but give an insight into the family of proteins and hence the functions of a protein of interest. The family IDs attached to the proteins of the RBP2GO database could thus give more information about their RNA-binding capacities. In turn, they were studied in the same fashion as the RBDs. In my initial lists of 809 literature-derived InterPro IDs and 2235 database-mined InterPro IDs, a total of 1028 of them were family IDs (Figure 12). These IDs were submitted to the same selection process as the one used for the RBDs, to select annotations enriched in the RBPs compared to the non-RBPs. In the end, 627 RNA-related family IDs, from now on called Rfam IDs, were selected (Figure 12).

Similar to the analysis of RBDs, the repartition of these IDs in the different species of the database was characterized. As was observed for the RBDs, the best studied species displayed a lower number of Rfam IDs present only in non-RBPs, while the number of Rfam IDs present only in RBPs was lower for the less studied species. Again, Ec and Pf appeared as exceptions (Figure 13A). For instance, Dr had less than 10% of its Rfam IDs in RBPs only, while 67% of Mm Rfam IDs could be found in RBPs only. This further underlines the lack of coverage in the less studied species, in which more Rfam-ID containing proteins were not detected as RBPs as in the better-studied species. Interestingly, the proportion of RBPs that did not contain any RBD but have an Rfam ID was low, even for well-studied species. In Hs, this amounted to 6.1% of the RBPs, and went up to 18.5% in ST although this species only has 226 RBPs in total. For the four most studied species, namely Hs, Mm, Sc and Dr, the proportion of RBPs lacking any RNA-related annotation from InterPro, meaning not annotated with an Rfam ID nor with an RBD, varied from 53% for Mm up to 70% for Hs. In contrast, only 23.8% of the RBPs in Pf did not exhibit any RNA-related annotation, although RBPs represented more than 18% of its proteome. Conversely, in lowly-studied species, the proportion of non-RBPs that had no RBD but had an Rfam ID was higher than for more studied species. This proportion amounted to 1.5% for Dr, when it represented less than 0.9% in all of the 4 more studied species. This again indicates that there is a lack of knowledge regarding RBPs in the species with a limited number of studies available.

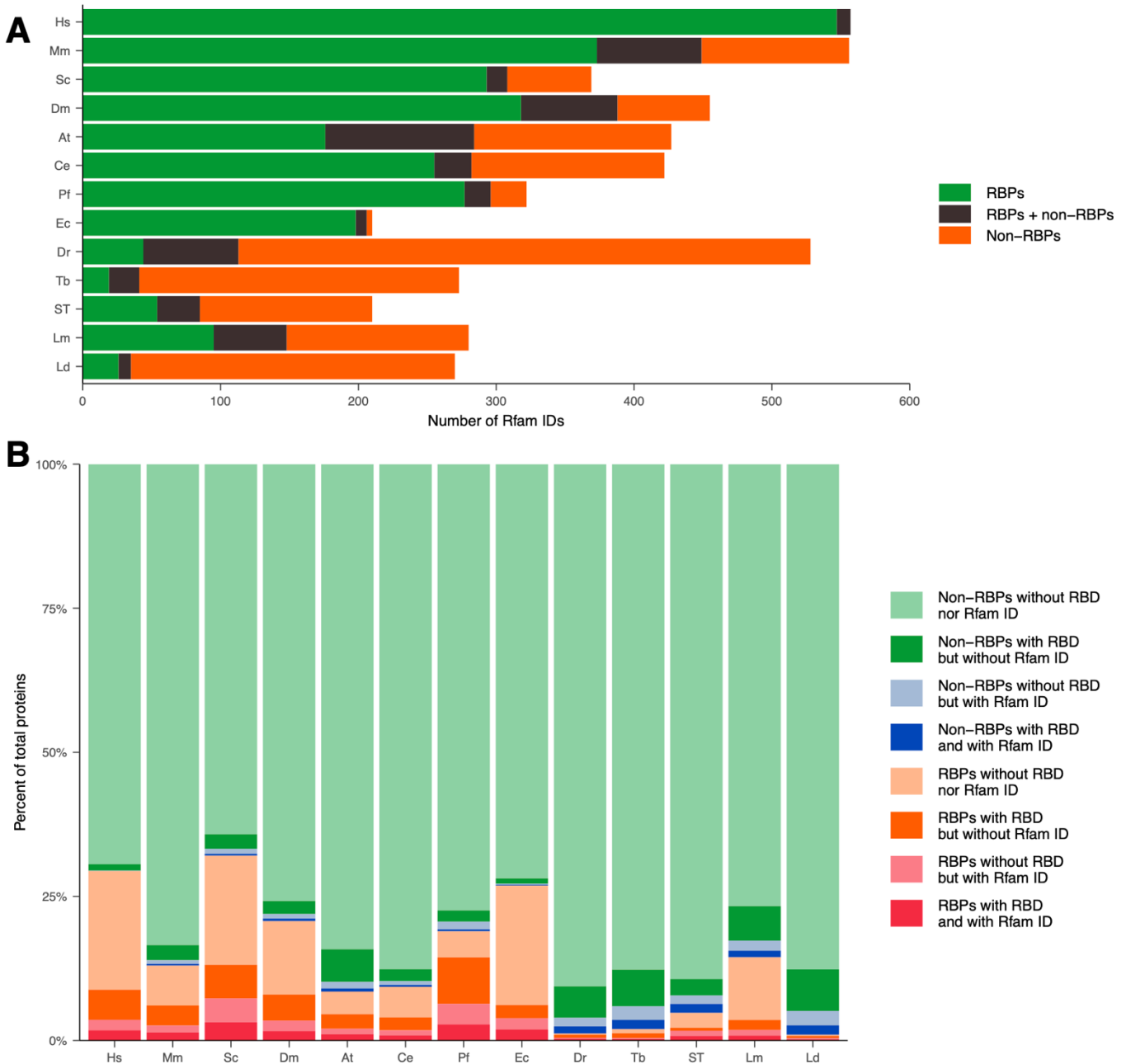


Figure 13: Repartition of the Rfam IDs in the species and the proteins of the RBP2GO database

A. Number of selected Rfam IDs per species grouped according to their distribution in RBDs present only in RBPs (green), RBDs present in both non-RBPs + RBPs (black) and RBDs present only in non-RBPs (orange). B. Proportion of the proteom in % of non-RBPs without RBD nor Rfam ID (light green), non-RBPs with RBD but without Rfam ID (dark green), non-RBPs without RBD but with Rfam ID (light blue), non-RBPs with RBD and Rfam ID (dark blue), 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 red) and RBPs with RBD and Rfam ID (dark red) in the different species.

2.5. Rfam IDs also help to discriminate strong RBP candidates

The relationship between the presence of an Rfam ID and the RBP2GO score was also studied to assess the information these annotations bring about the RNA-binding ability of a protein. When comparing the RBP2GO score for proteins annotated or not with an Rfam ID, the proteins with an Rfam ID showed a significantly higher score in all species (Figure 14A). Again, the proteins were split into four groups: RBPs and non-RBPs with or without an Rfam ID. Regarding non-RBPs, all species except Hs exhibited a higher RBP2GO score for non-RBPs with an Rfam ID compared to the non-RBPs without an Rfam ID. The lack of significance in Hs could be explained by the low number of non-RBPs with an Rfam (14 proteins). The same was observed in RBPs, with a significantly higher score for proteins with an Rfam ID. Here, Tb was an exception, again certainly due to the low number of RBPs with an Rfam ID in this species (36 proteins in total) (Figure 14B).

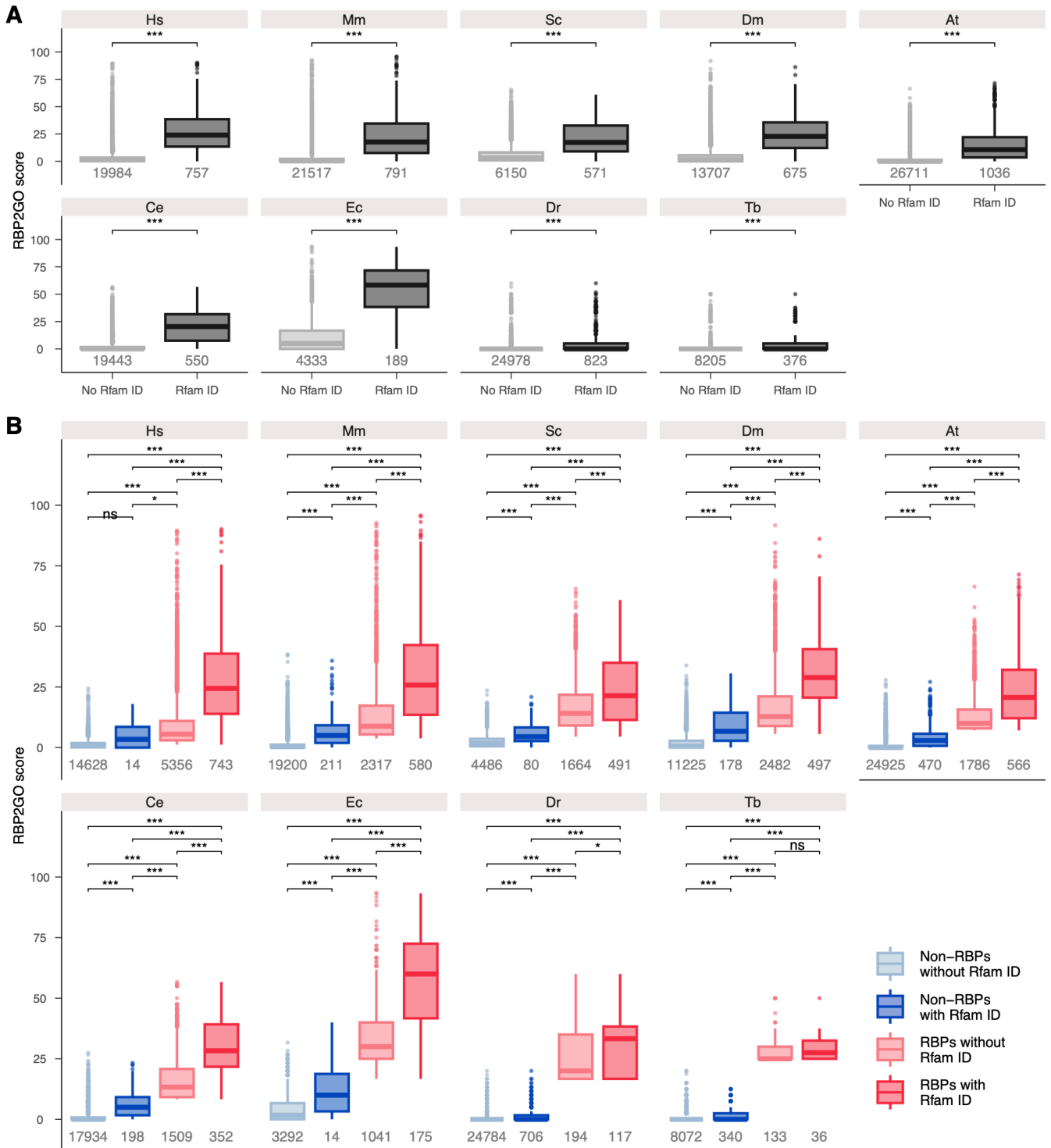


Figure 14: Influence of the presence of Rfam IDs on the RBP2GO score

A. Boxplot representing the distribution of the RBP2GO score in proteins without Rfam ID (no Rfam ID, light grey) and with Rfam ID (Rfam ID, dark grey). B. Same as in A., but separated into non-RBPs without Rfam ID (light blue), non-RBPs with Rfam ID (dark blue), RBPs without Rfam ID (light red) and RBPs with Rfam ID (dark red). *, **, *** and **** correspond to p-values < 0.05, 0.01, 0.001 and 0.0001, respectively, resulting from a Wilcoxon rank sum test.

The relationship between the proportion of proteins annotated with Rfam IDs and the RBP2GO score was also analyzed. We can see that the same positive correlation up to a score of 50 could be observed in all species, as for RBDs, with the same exceptions of Dr and Tb (Figure 15). But in contrast to the RBDs, no plateau could be observed after 50; it does not seem that proteins that have a high RBP2GO score were automatically annotated with an Rfam ID.

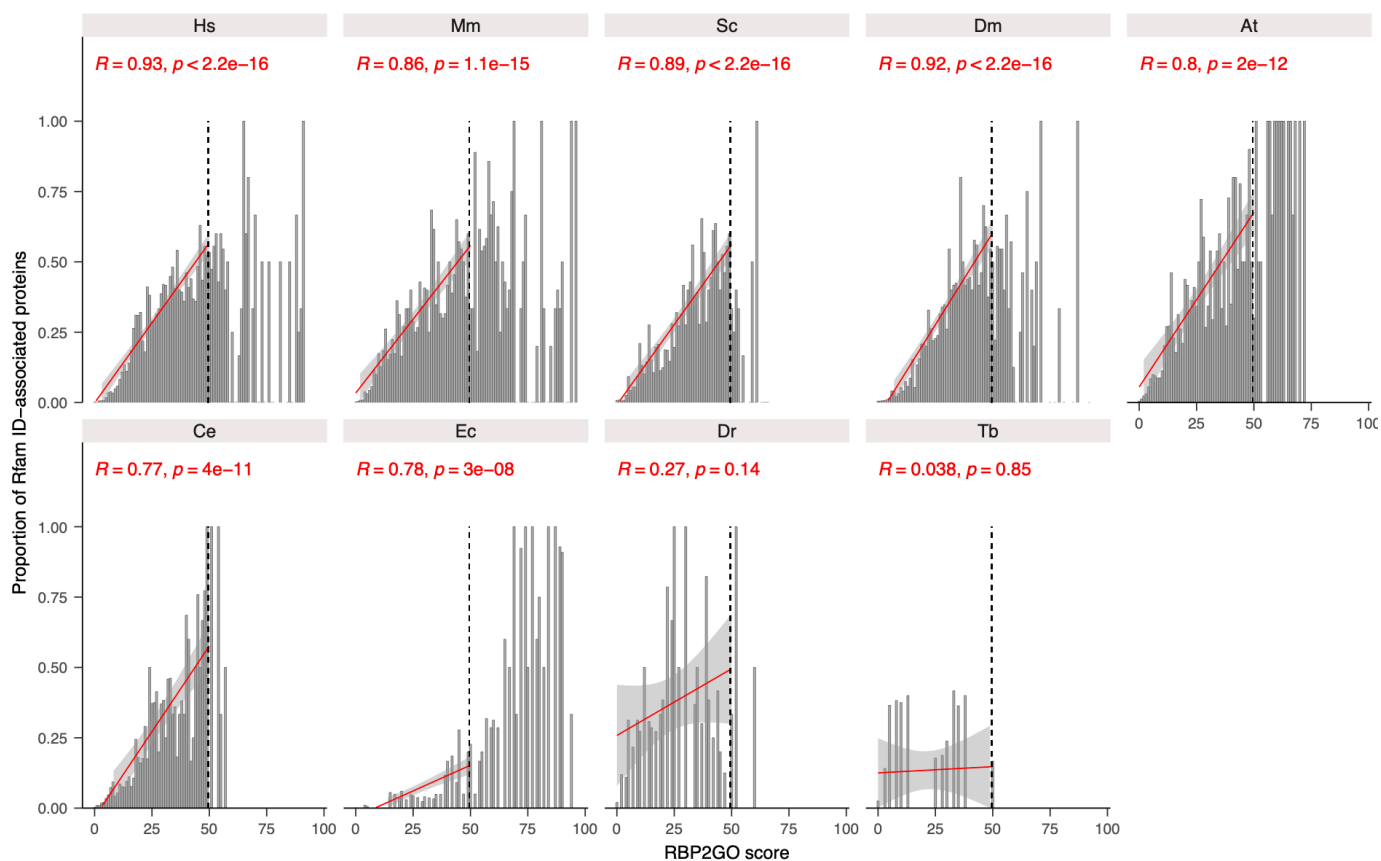


Figure 15: Proportion of Rfam ID-associated proteins (in %) for each unit of the RBP2GO score

The red line represents a linear regression between 0 and 50. R represents the Person's correlation coefficient.

Finally, the proteins were separated into eight different groups to assess the different effects of the presence of an RBD and/or an Rfam ID on the RBP2GO scores of the proteins. The proteins were grouped as follows: non-RBPs with an RBD, an Rfam ID, both or none, and RBPs with an RBD, an Rfam ID, both or none (Figure 16). For the non-RBPs, proteins with an RBD or an Rfam ID displayed a significantly higher score in all species compared to the proteins with neither annotation. Comparing the non-RBPs with only an RBD to the non-RBPs with only an Rfam ID yielded a more heterogeneous result in different species: while the non-RBPs with an Rfam ID showed a significantly higher score in some species (Mm, At, Ce, Dr), other species did not show a significant difference (Hs, Sc, Dm, Ec) or even displayed a significantly higher score in RBD-containing non-RBPs (Tb). Similarly, the non-RBPs with both annotations, RBD and Rfam ID, had a significantly higher score than the non-RBPs with only one type of annotation in Mm and Dm. In At, Ce, Dr, and Tb, the non-RBPs with both annotations had a significantly higher score than non-RBPs with only an RBD.

The same conclusions could be drawn for RBPs. The RBPs with an RBD or an Rfam ID possessed a significantly higher score in all species compared to the RBPs with no annotation. The difference between RBPs either with only an RBD or with only an Rfam ID again was heterogeneous: some species showed a significantly higher score for the RBPs with an Rfam ID (Hs, Dm, At, Ce, Ec), others displayed no statistically significant difference (Mm, Dr, Tb) or even a significantly lower score for RBPs with an Rfam IDs (Sc). Finally, the RBPs with the two types of annotations had a significantly higher score than the RBPs with either only an RBD or an Rfam ID in Hs, Mm, Dm, Ce and Ec. In Sc, the RBPs with both annotations had a significantly higher score than RBPs with only an Rfam ID, while the opposite was the case in Tb (but based on very few proteins). In At, the RBPs with both annotations had a significantly higher score than RBPs with only an RBD.

As a result, Rfam IDs also correlated with the RNA-binding capacities of a protein similar to what has been shown for the RBDs. However, they do not allow to annotate more RBPs than the RBDs; two thirds of the human RBPs are still left without an RNA-related annotation from InterPro.

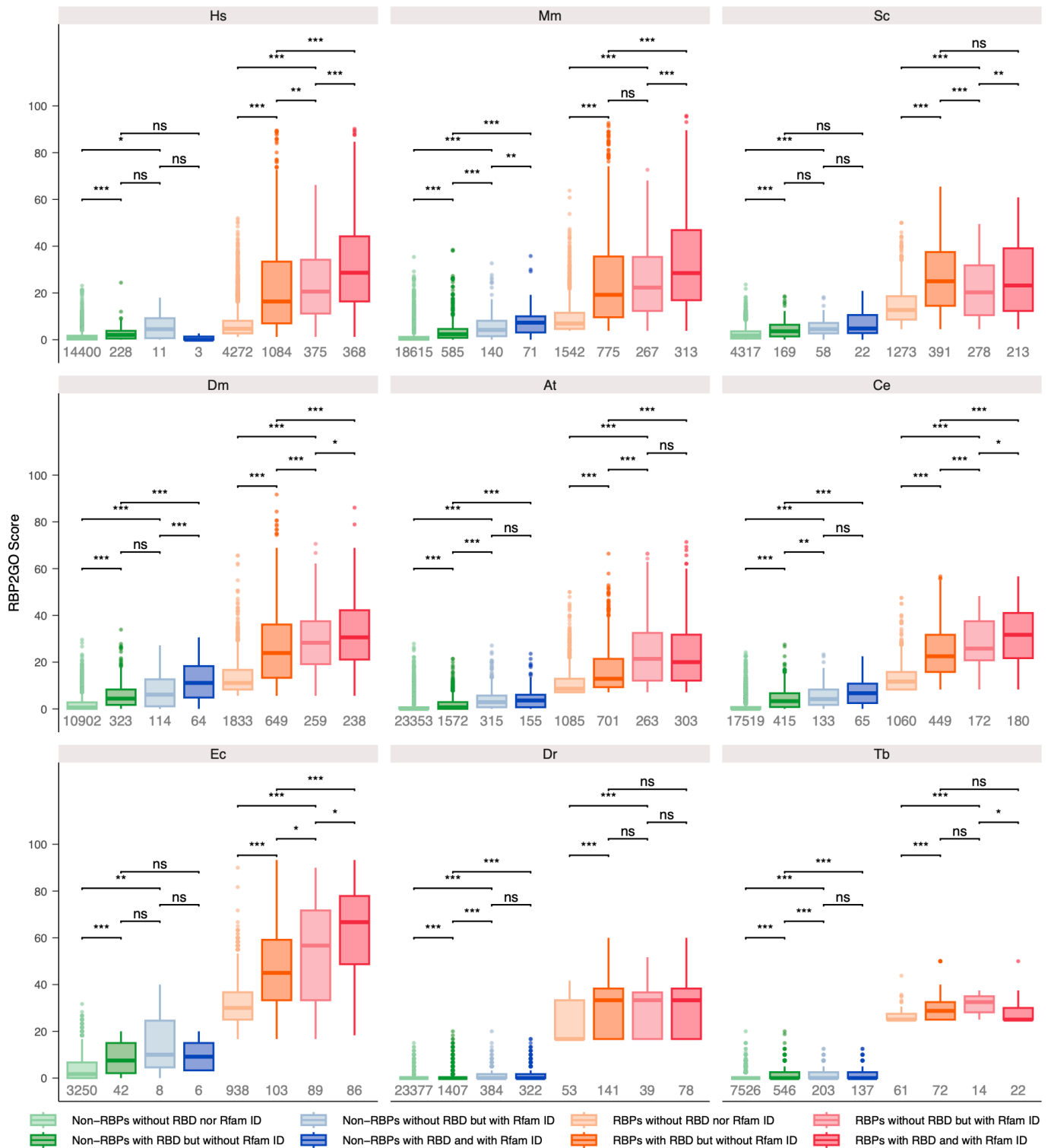


Figure 16: Influence of the presence of Rfam IDs in combination with RBDs on the RBP2GO score

Boxplot representing the distribution of the RBP2GO score in non-RBPs without RBD nor Rfam ID (light green), non-RBPs with RBD but without Rfam ID (dark green), non-RBPs without RBD but with Rfam ID (light blue), non-RBPs with RBD and Rfam ID (dark blue), 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 red) and RBPs with RBD and Rfam ID (dark red). *, **, *** and **** correspond to p-values < 0.05, 0.01, 0.001 and 0.0001, respectively, resulting from a Wilcoxon rank sum test.

2.6. Distribution of disordered regions in the proteins of the RBP2GO database

Taking into consideration that 70% of human RBPs do not contain an RBD nor are they annotated with an Rfam ID, I decided to investigate the presence of intrinsically disordered regions, or IDRs, in the proteins of the database. As was previously stated, IDRs can participate in RNA-binding, through binding to RNA on their own or acting in conjunction with canonical RBDs.

I chose to use the prediction from the MobiDB-lite algorithm, as its results come from the compilation of several different algorithms as well as experimental data from the PDB database (142, 152). These predictions are also present on the Uniprot and InterPro databases (144, 145), and the MobiDB-Lite algorithm is considered more conservative, so less prone to false positives (147). The coordinates of the IDRs for each protein as well as the disordered content fraction were downloaded from the MobiDB website (143). When looking at the distribution of RBPs and non-RBPs with or without an RBD, the RBPs with an RBD contained more disordered regions than the RBPs without an RBD based on their minimum disordered fraction (fraction of the protein length covered by an IDR) except for Tb (Figure 17A). For example, 32.3% of the human RBPs with an RBD had a minimum disordered fraction of 25%, while this was the case for only 20.5% of the human RBPs without an RBD. In Hs, Mm, Sc, Dm, Ce and Dr, the non-RBPs with an RBD also contained more disordered regions than the non-RBPs without an RBD (Figure 17A). This result is confirmed by the statistical comparison of the disordered fraction per protein in each of the four groups. For all species, except Tb and Ld due to their low number of RBPs, the RBPs with an RBD showed a significantly higher disordered fraction than the RBPs with no RBD. The same difference could be observed between non-RBPs with and without an RBD except in Pf, Lm and Ld (Figure 17B). These results show that IDRs are generally enriched in proteins already having an RBD compared to the proteins that do not contain any RBD.

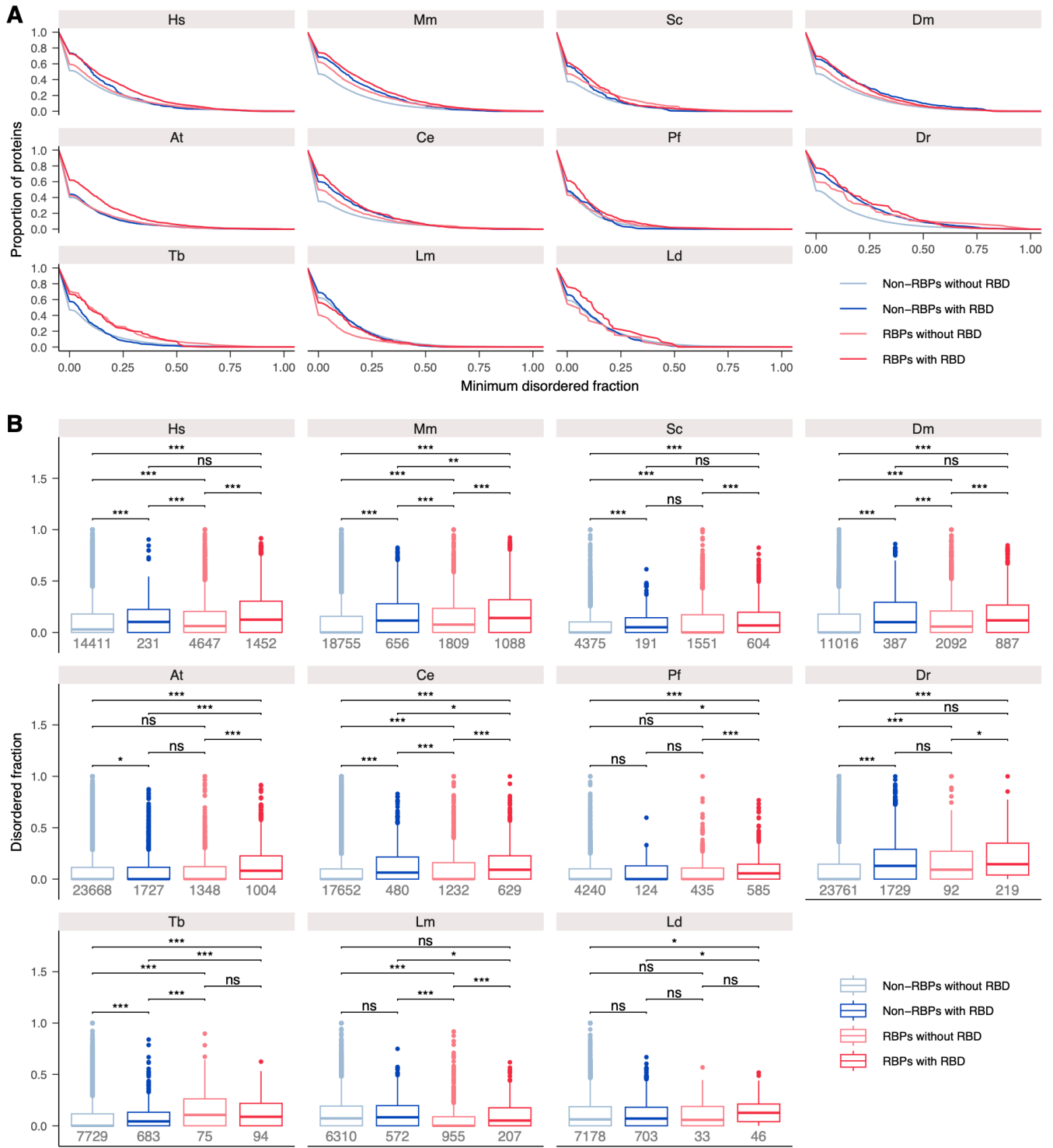


Figure 17: Distribution of the disordered regions in the proteins of the RBP2GO database
 A. Cumulative proportion of proteins against their minimum disordered fraction, in the four groups of proteins: non-RBPs without RBD (light blue), non-RBPs with RBD (dark blue), RBPs without RBD (light red) and RBPs with RBD (dark red). B. Boxplot showing the distribution of the disordered fraction in the four groups of proteins: non-RBPs without RBD (light blue), non-RBPs with RBD (dark blue), RBPs without RBD (light red) and RBPs with RBD (dark red). The proteins containing no disordered region were also included. The numbers given in grey are the numbers of proteins in each group. *, **, *** and **** correspond to p-values < 0.05, 0.01, 0.001 and 0.0001, respectively, resulting from a Wilcoxon rank sum test.

The MobiDB database also categorizes different types of disorder (143). Since the types of disordered regions known to bind to RNA are known to be enriched in charged amino acids (127), I repeated the previous analysis, taking into account only the “polyampholyte” IDRs (153), meaning the IDRs enriched in both positively or negatively charged amino acids. Here, the same result could be observed as previously: all species exhibited a higher proportion of RBPs with an RBD containing a given fraction of polyampholyte IDRs compared to the RBPs without an RBD except for Tb (Figure 18A). The same difference between the non-RBPs with an RBD and the non-RBPs without an RBD can be observed in Hs, Mm, Sc, Dm, At, Ce, Dr and Tb. Again, the RBPs containing an RBD had a significantly higher polyampholyte disordered fraction in all species except Tb, Dr and Ld, likely due to the small group sizes (Figure 18B). Also, the difference observed between the two non-RBP groups are confirmed in the most species (Hs, Mm, Sc, Dm, At, Ce, Dr, Tb) with a significantly higher polyampholyte disordered fraction for the proteins with an RBD (Figure 18B).

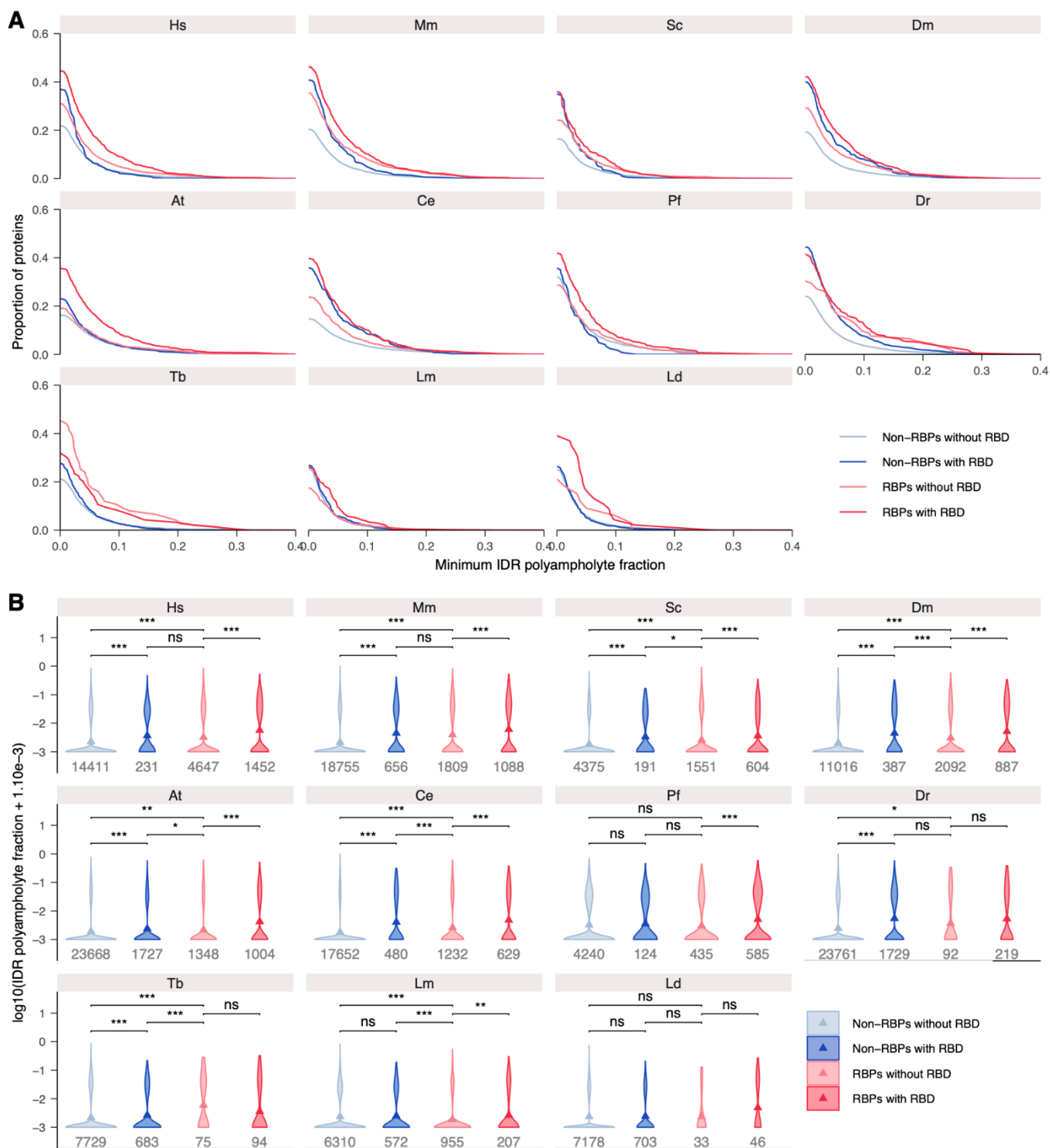


Figure 18: Distribution of the polyampholyte IDRs in the proteins of the RBP2GO database
 A. Cumulative proportion of proteins against their minimum disordered fraction of polyampholyte IDRs, in the four groups of proteins: non-RBPs without RBD (light blue), non-RBPs with RBD (dark blue), RBPs without RBD (light red) and RBPs with RBD (dark red). B. Boxplot showing the distribution of the disordered fractions of polyampholyte IDRs in the four groups of proteins: non-RBPs without RBD (light blue), non-RBPs with RBD (dark blue), RBPs without RBD (light red) and RBPs with RBD (dark red). The proteins containing no disordered region were also included. The numbers given in grey are the numbers of proteins in each group. *, **, *** and **** correspond to p-values < 0.05, 0.01, 0.001 and 0.0001, respectively, resulting from a Wilcoxon rank sum test.

Finally, I also studied the presence of coiled-coil regions in the same four groups of proteins, as these regions have also been found to be enriched in RBPs (154). The results are presented here only for Hs, Mm, Sc and Dm, since no visible difference could be observed in the other species. Contrary to the other types of disorder, coiled-coil regions were more present in the proteins with no RBD than in the RBD-containing proteins, both in RBPs and non-RBPs, in all four species (Figure 19A). Furthermore, the group with the highest proportion of proteins with coiled-coil regions were the RBPs with no RBD. The higher proportion of coiled-coil regions in RBPs with no RBD compared to RBPs with RBDs, as well as in non-RBPs with no RBD compared to non-RBPs with RBD was found to be statistically significant in Hs, Mm and Dm. However, no significant difference could be found between the non-RBPs and the RBPs with no RBD (Figure 19B).

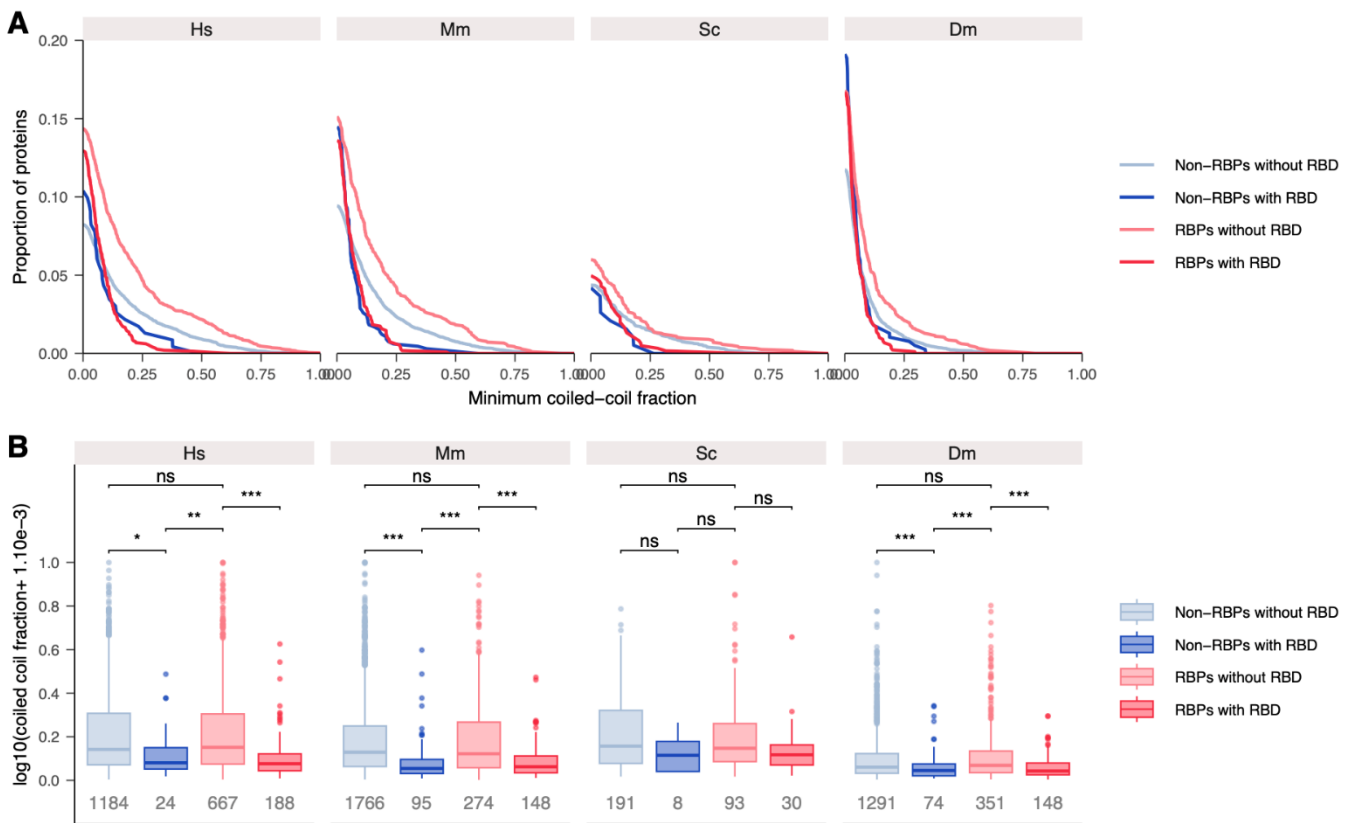


Figure 19: Distribution of the coiled-coil regions in Hs, Mm, Sc and Dm

A. Cumulative proportion of proteins against their minimum coiled-coil fraction, in the four groups of proteins: non-RBPs without RBD (light blue), non-RBPs with RBD (dark blue), RBPs without RBD (light red) and RBPs with RBD (dark red). **B.** Boxplot showing the distribution of the coiled-coil fraction in the four groups of proteins: non-RBPs without RBD (light blue), non-RBPs with RBD (dark blue), RBPs without RBD (light red) and RBPs with RBD (dark red). The proteins containing no coiled-coil region were also included. The numbers given in grey are the numbers of proteins in each group. *, **, *** and **** correspond to p-values < 0.05, 0.01, 0.001 and 0.0001, respectively, resulting from a Wilcoxon rank sum test.

To evaluate the impact of the presence of an IDR on the RBP2GO score of a protein, the proteins were once again divided into 8 groups: RBPs with no domain, an RBD, an IDR, or both, and non- with no domain, an RBD, an IDR, or both or (Figure 20). The RBP2GO scores of these groups were then compared. Non-RBPs with an IDR had a significantly higher score than non-RBPs with no domain in most of the species, except Ec and Tb, while the RBPs with an IDR had a significantly higher score than the RBPs with no domain in only Hs and Ec. In Sc and Ce, the RBPs with an IDR even had a significantly lower score than the RBPs with no domain. In the other five species, no significant difference could be observed. Additionally, no difference was present between the non-RBPs with an RBD only and the non-RBPs with both an RBD and an IDR, except for At and Sc. Moreover Hs, Mm, Sc and Ec showed a higher RBP2GO score for RBPs with both domains compared to RBPs with only an RBD.

Furthermore, no correlation could be found between the disordered fraction of a protein and its RBP2GO score (Figure 21A). In Hs, Mm, Dm, At, Ec and Dr, a positive linear relationship could be seen between the score of RBPs and their disordered content fraction, but none of the correlation coefficient was higher than 0.18, underlining a very low correlation. No correlation at all could be observed for the non-RBPs.

Finally, I compared the disordered fraction of the top 10% of RBPs with no RBD classified by RBP2GO score (high score) with the disordered fraction of the bottom 10% RBPs with no RBD (low score, Figure 21B). The disordered fraction of the proteins with a high score was significantly higher than the fraction of proteins with a low score in only Hs and Mm, and was even found to be significantly lower in Ce. In all other proteins, no difference could be observed. Therefore, human and mouse RBPs that do not have any RBD but a high score exhibited a higher disordered fraction, which might explain how they bind RNA, but this could not be extended to the other species of this study.

In the end, the presence of a disordered region in a protein does not show a strong correlation regarding the candidature of a protein as an RBP.

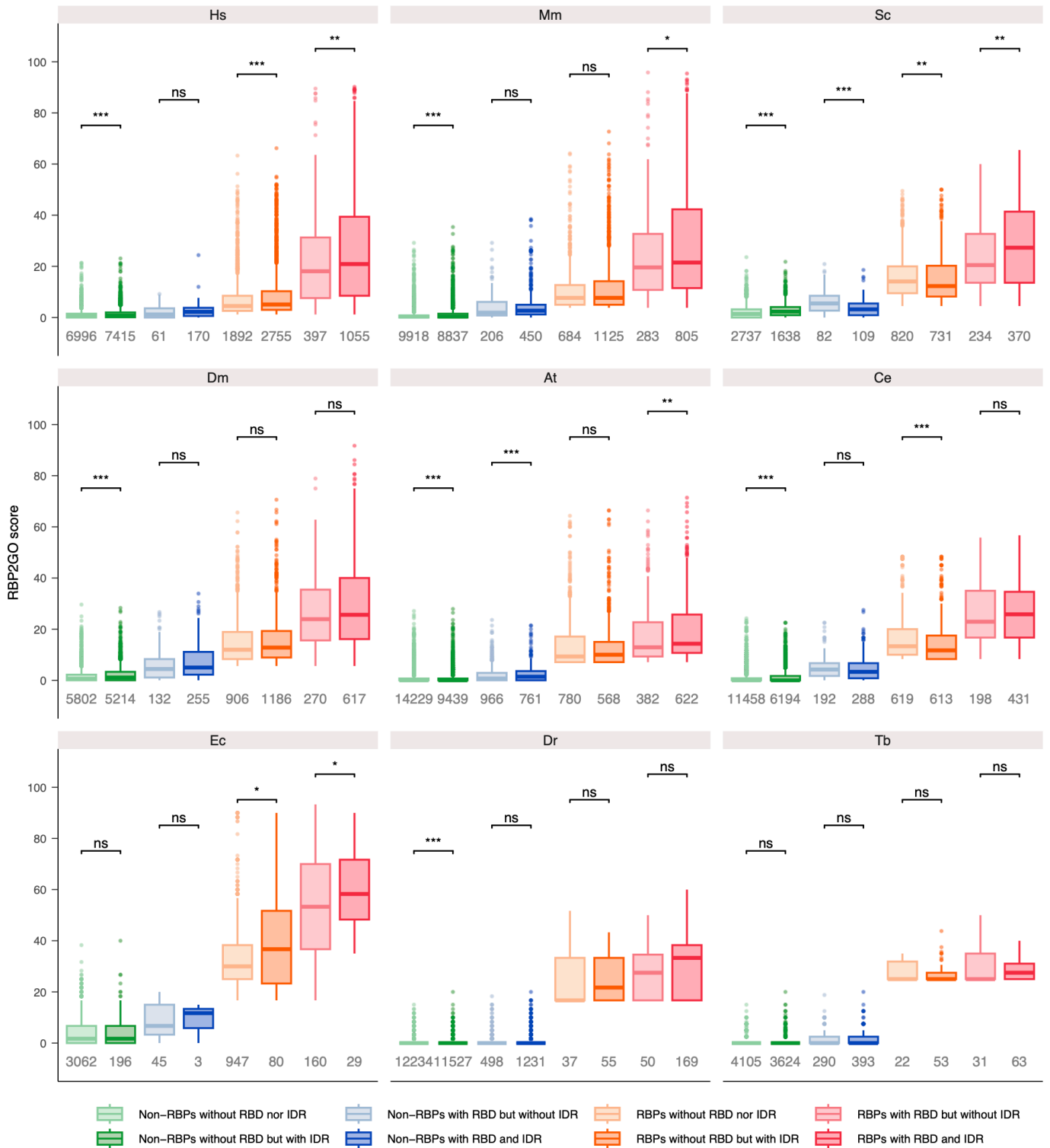


Figure 20: Influence of the presence of an IDR on the RBP2GO score

Boxplot representing the distribution of the RBP2GO score in non-RBPs without RBD nor IDR (light green), non-RBPs without RBD but with IDR (dark green), non-RBPs with RBD but without IDR (light blue), non-RBPs with RBD and IDR (dark blue), RBPs without RBD nor IDR (light orange), RBPs without RBD but with IDR (dark orange), RBPs with RBD but without IDR (light red) and RBPs with RBD IDR dark red). *, **, *** and **** correspond to p-values < 0.05, 0.01, 0.001 and 0.0001, respectively, resulting from a Wilcoxon rank sum test.

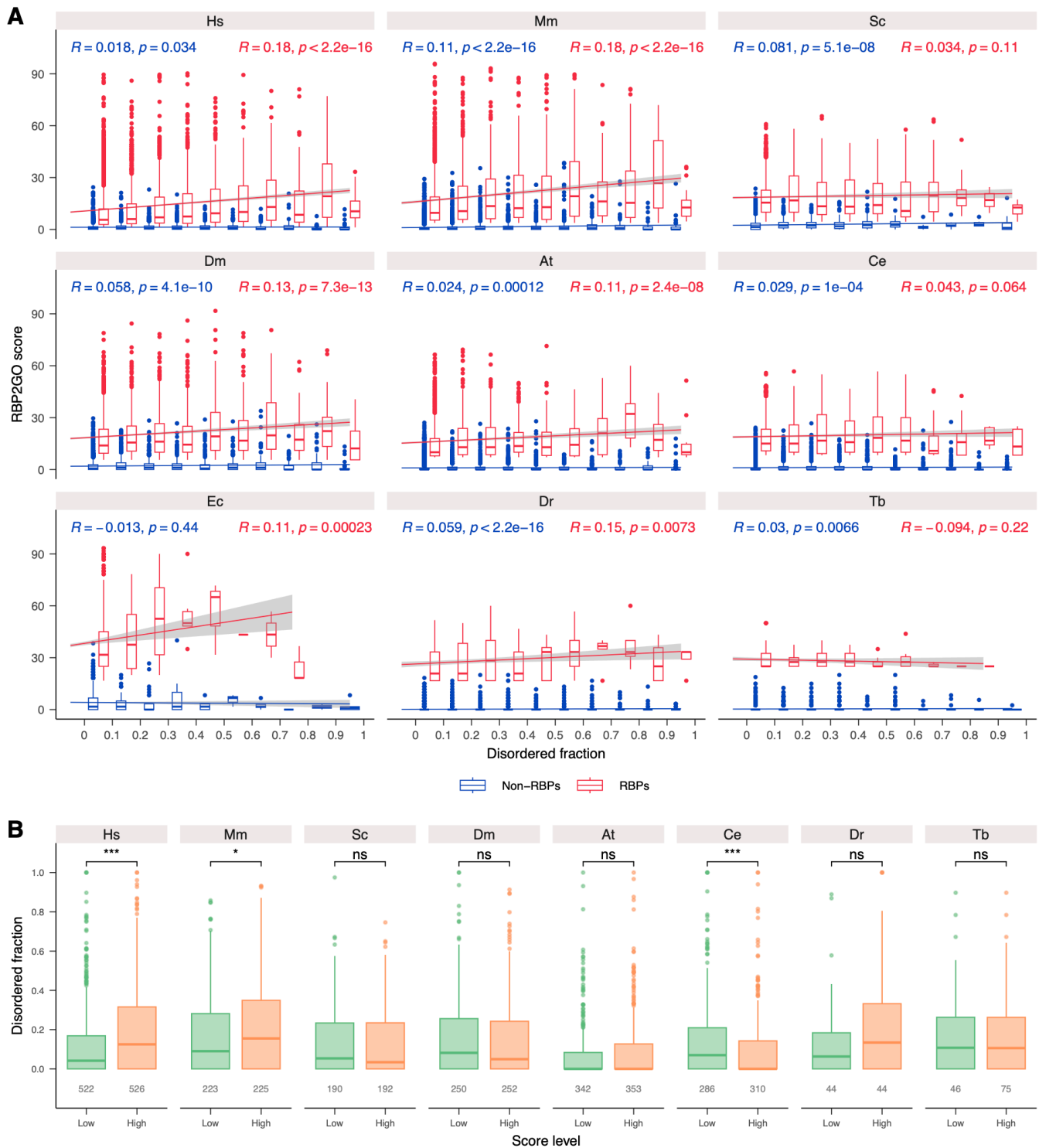


Figure 21: Correlation between the presence of disordered regions and the RBP2GO score
 A. Boxplot representing the distribution of the RBP2GO score for non-RBPs (blue) and RBPs (red) in function of the disordered fraction of the proteins. R represents the Pearson's correlation coefficient. B. Boxplot showing the distribution of the disordered fraction in the bottom 10% of the RBPs with no RBD and with the lowest RBP2GO score (green) and in the top 10% of the RBPs with no RBD and the highest RBP2GO score (orange). *, **, *** and **** correspond to p-values < 0.05, 0.01, 0.001 and 0.0001, respectively, resulting from a Wilcoxon rank sum test.

2.7. Prediction of new RBP candidates

Regarding the lack of RBP detection in the less studied species as well as the predictive potential of RBDs that was highlighted in the analyses above, I explored the idea that non-RBPs with RBDs could represent new RBP candidates. I compared the number of these proteins with the number of already detected RBPs in species with a high number of datasets compiled in RBP2GO and species with a low number of datasets. It appeared clearly that taking into account non-RBPs with RBDs could almost double the pool of RBP candidates for the less studied species. It would add 6166 to the 8419 RBPs present in RBP2GO, while the 1463 non-RBPs with RBD in well-studied species would yield only 10% more RBPs (Figure 22A).

To further investigate the potential of these proteins, I performed two GO term enrichment analyses using Panther (155) in the ten species that were available in this database. I then selected the terms enriched in at least five species, with a minimum average fold enrichment of four. Most of the enriched GO molecular function terms were related to RNA and DNA (in green), metabolism (in blue), or chromatin regulation (in orange, Figure 22B). Regarding the GO biological process terms, a large number of the most enriched terms were also related to RNA or DNA (in green), and some of them to chromatin regulation (in orange), while only one term appeared linked to metabolism (in blue, Figure 22C). This confirms that non-RBPs containing RBDs are enriched in functions related to RNA or in processes involving known RBPs.

Finally, I selected non-RBPs with RBDs and a high score in Mm and performed a literature search to explore whether some of these proteins have already been shown to bind to RNA. I found seven proteins with different RBDs for which individual studies demonstrated their ability to bind RNA (Table 2). These proteins contained different types of RBDs: RRM (IPR000504, IPR003954), zinc fingers (IPR000571, IPR001878), a poly(a) polymerase domain (IPR002058), a SAP domain (IPR003034) and an exonuclease domain (IPR013520). The proteins harbored only one type of domain, like the Rbmy proteins, or two different types of RBDs like Tut7. Furthermore, these proteins did not belong to the same families, and could be divided in three groups: the proteins with an RRM, the proteins involved in splicing and the RNA-modifying enzymes. This literature search thus showed that unrelated proteins with different RBDs had already been shown to bind to RNA. These results confirmed that non-RBPs containing RBDs represent a pool of promising RBP candidates.

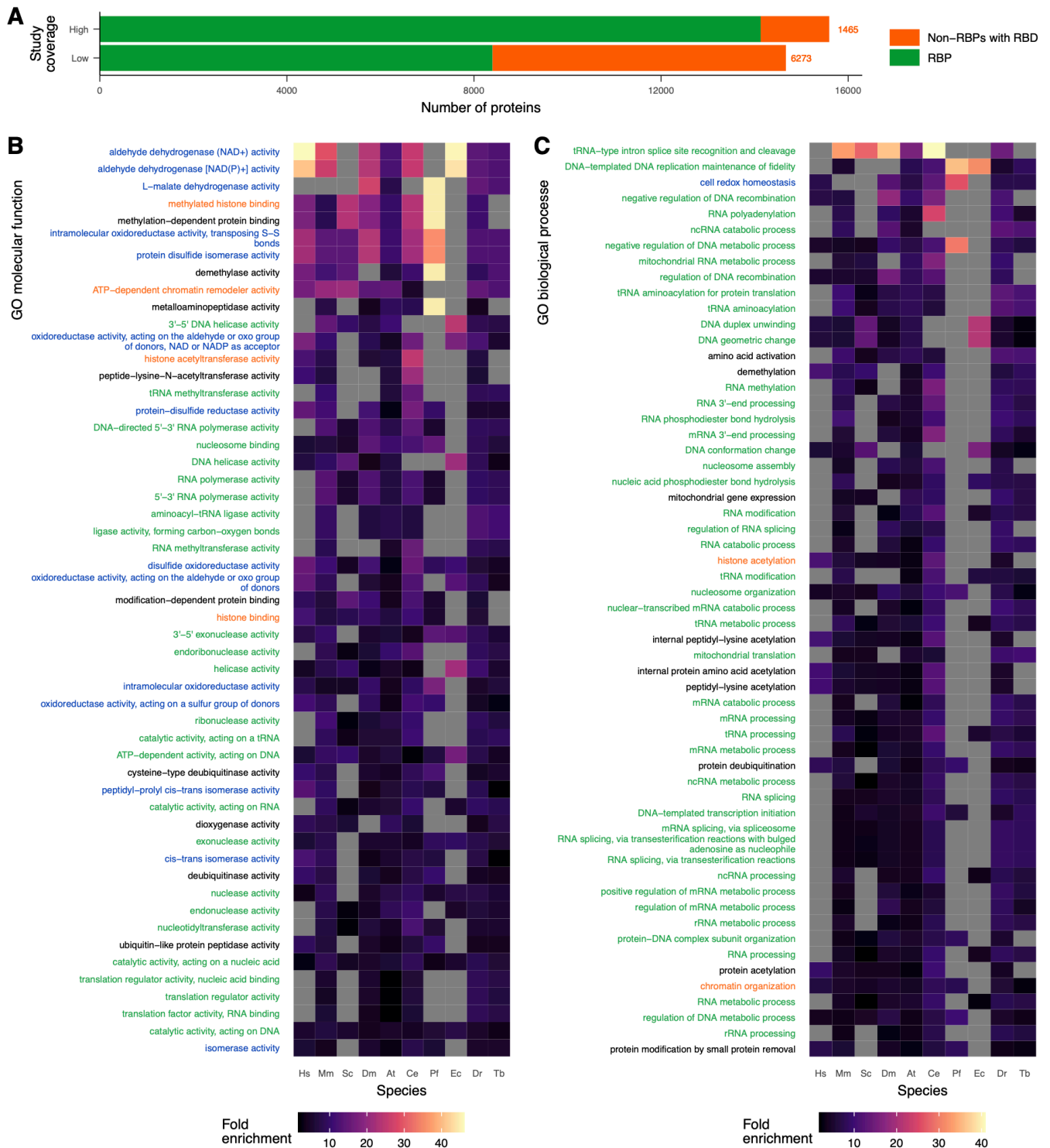


Figure 22: Discovery of new RBP candidates using the RBD content of the non-RBPs

A. Number of RBPs and non-RBPs with an RBD for species with a high (≥ 9) or a low (< 9) number of studies. B. Heatmap showing the most enriched molecular function GO terms in the non-RBPs with an RBD compared to the proteome of each species (for more details, see the Materials and Methods section). The GO terms were classified by mean enrichment, from top to bottom. The terms in blue are related to metabolism, the ones in green to DNA and RNA, and the ones in orange to chromatin regulation. C. Same as in B., for the biological process GO terms.

Uniprot ID	Gene name	Protein name	List of RBDs	RBP2GO score	Experiment	Ref
O35698	Rbmy1a1	RNA-binding motif protein, Y chromosome, family 1 member A1	IPR000504	38.1	PR-CLIP	(147)
Q60990	Rbmy1b	RNA-binding motif protein, Y chromosome, family 1 member B	IPR000504	38.1	PR-CLIP	(147)
E9Q6E5	Srsf11	Serine and arginine-rich-splicing factor 11	IPR000504	35.8	EMSA, iclip	(148)
Q62377	Zrsr2	U2 small nuclear ribonucleoprotein auxiliary factor 35 kDa subunit-related protein 2	IPR000504 IPR003954 IPR000571	30	CLIP	(149)
Q3US41	Esrp1	Epithelial splicing regulatory protein 1	IPR000504	26.5	EMSA - proof only in human	(150) – only in human cell lines
Q5BLK4	Tut7	Terminal uridylyltransferase 7	IPR001878 IPR002058	21.2	RNA-binding assessed by similarity with the human protein, but uridylyltransferase activity showed in mouse	(151)
Q7TMF2	Eri1	3'-5' exoribonuclease 1	IPR003034 IPR013520	19.6	Co-IP of rRNA in Eri1 tandem-affinity purification	(152)

Table 2: New RBP candidates validated by a literature search

2.8. Prediction of new RBD candidates in proteins detected as RBPs but lacking known RBDs

Considering that the RBPs with no RBD did not show an enrichment in disordered regions, I hypothesized that these proteins may contain unknown RBDs that were not part of the original list from which the RBDs were selected. An enrichment analysis was thus performed on human proteins to identify these domains. From the RBPs with no RBDs, the top 20% according to their RBP2GO score were selected as the dataset of interest, and three reference datasets were constituted: the RBPs with no RBDs from the bottom 20% according to their RBP2GO score, from now on referred to as “RBPs with a low score”, the non-RBPs with no RBD, and the whole human proteome. For this part of the study, only the “domain” and “repeat” types of InterPro IDs were considered. The enrichment analysis was performed in the dataset of interest against all three datasets of reference separately, and the significantly enriched domains were visualized (Figure 23A). Some of the significantly enriched domains were present in only a few proteins, so it was decided to select enriched domains that were present in at least 5 proteins in the dataset of reference. In total, 15 domains were found to be enriched in the top 20% scoring RBPs with no RBD; three of them were enriched compared to the RBPs with a low score, 13 compared to the whole proteome and all of them compared to the non-RBPs with no RBD (Figure 23B). Several of the domains are involved in cytoskeleton-binding, others are found in chaperones and a last group in importins. These 15 domains enriched in RBPs without a known RBD are from now on called “new RBDs”.

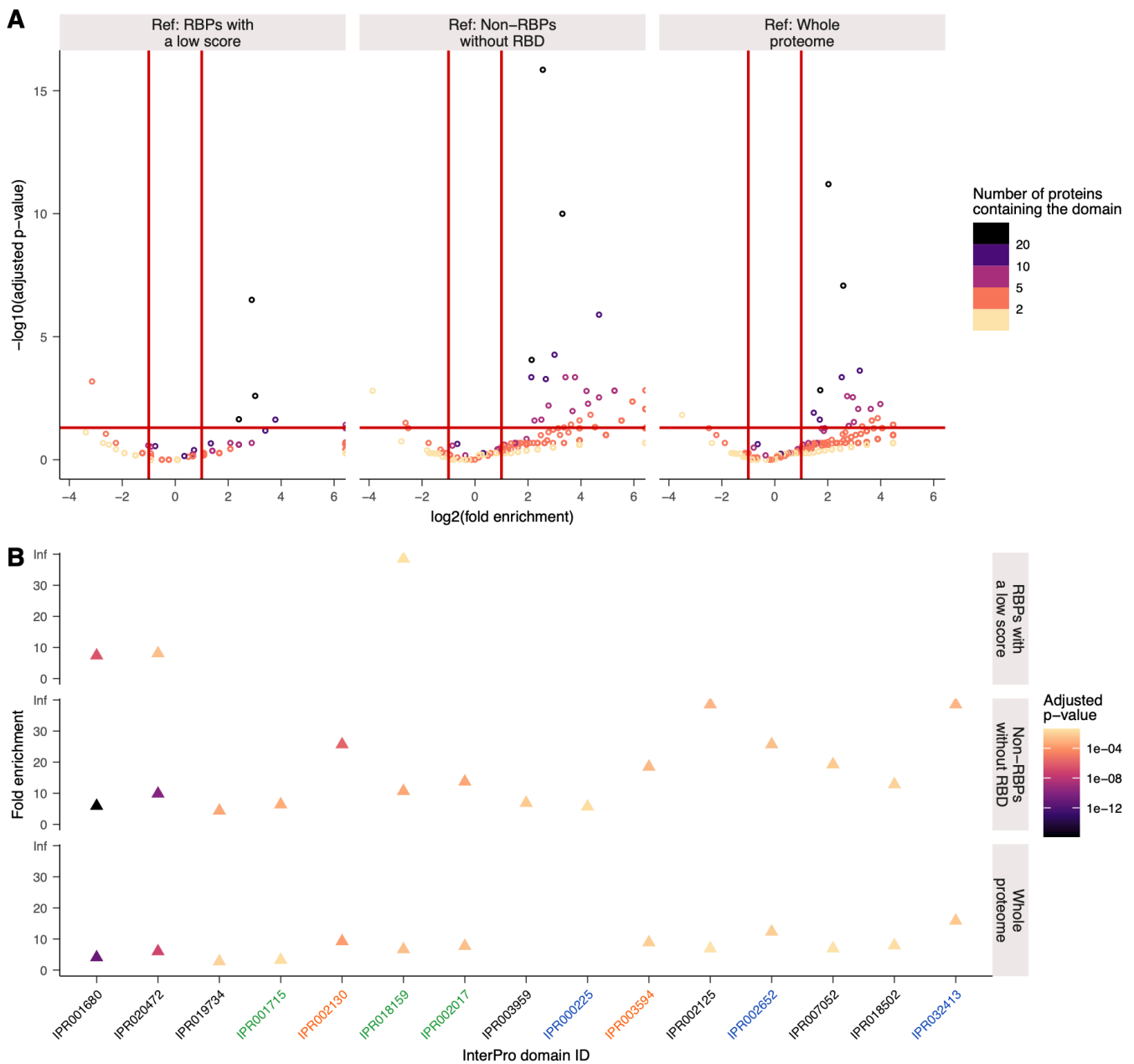


Figure 23: Discovery of new RBD candidates in RBPs with no RBD

A. Volcano plots showing the enrichment and its adjusted P-values for each of the InterPro domains present in the RBPs with no RBD and with the top 20% RBP2GO scores compared to three different control groups (RBPs with no RBD and bottom 20% RBP2GO scores / non-RBPs with no RBD / whole proteome). The *P*-values were calculated with a Fisher's exact test, and adjusted with the FDR method. (B) Same enrichment and P-values for the 18 selected InterPro IDs. For the selection process, see the Methods. The InterPro IDs in orange are related to chaperones, the ones in green to cytoskeletal protein-binding, and the ones in blue to importins.

To estimate the potential of these RBD candidates, I evaluated the proportion of RBPs that did not have any selected RBDs but contained one of these 15 new RBD candidates in Hs, but also in Mm, Sc and Dm (Figure 24A). In Hs, 5.5% of the RBPs harbored a new RBD versus only 1.9% of the non-RBPs. This expanded the pool of RBPs with an RBD by 23.5%, but also almost doubled the amount of non-RBPs with an RBD, although it still represented less than 2.5% of the human proteome. This higher proportion of proteins with a new RBD in RBPs compared to the non-RBPs was also observed in the three other well-studied species. For example, 5.9% of the RBPs in Mm contained only the new RBDs versus 2.2% of the non-RBPs. This enrichment was comprised between 2.2 and 3.2 for all species, and was also significant for all of them, even though the new RBDs were selected in human proteins only (Figure 24B). Finally, a comparison of the RBP2GO score for the proteins with a selected RBD, a new RBD or neither in the four species showed that, except in Mm, the proteins with the new RBDs had a significantly higher score than the proteins with no RBD (Figure 24C). However, this group of proteins still exhibited a significantly lower RBP2GO score than the proteins with selected RBDs in all species.

To summarize, an enrichment analysis in the domains present in RBPs with no RBD and a high score was able to detect 15 new candidate RBDs. These domains are enriched in all four most studied species of the RBP2GO database and the proteins containing them have a higher RBP2GO score than the proteins with no selected nor new RBD.

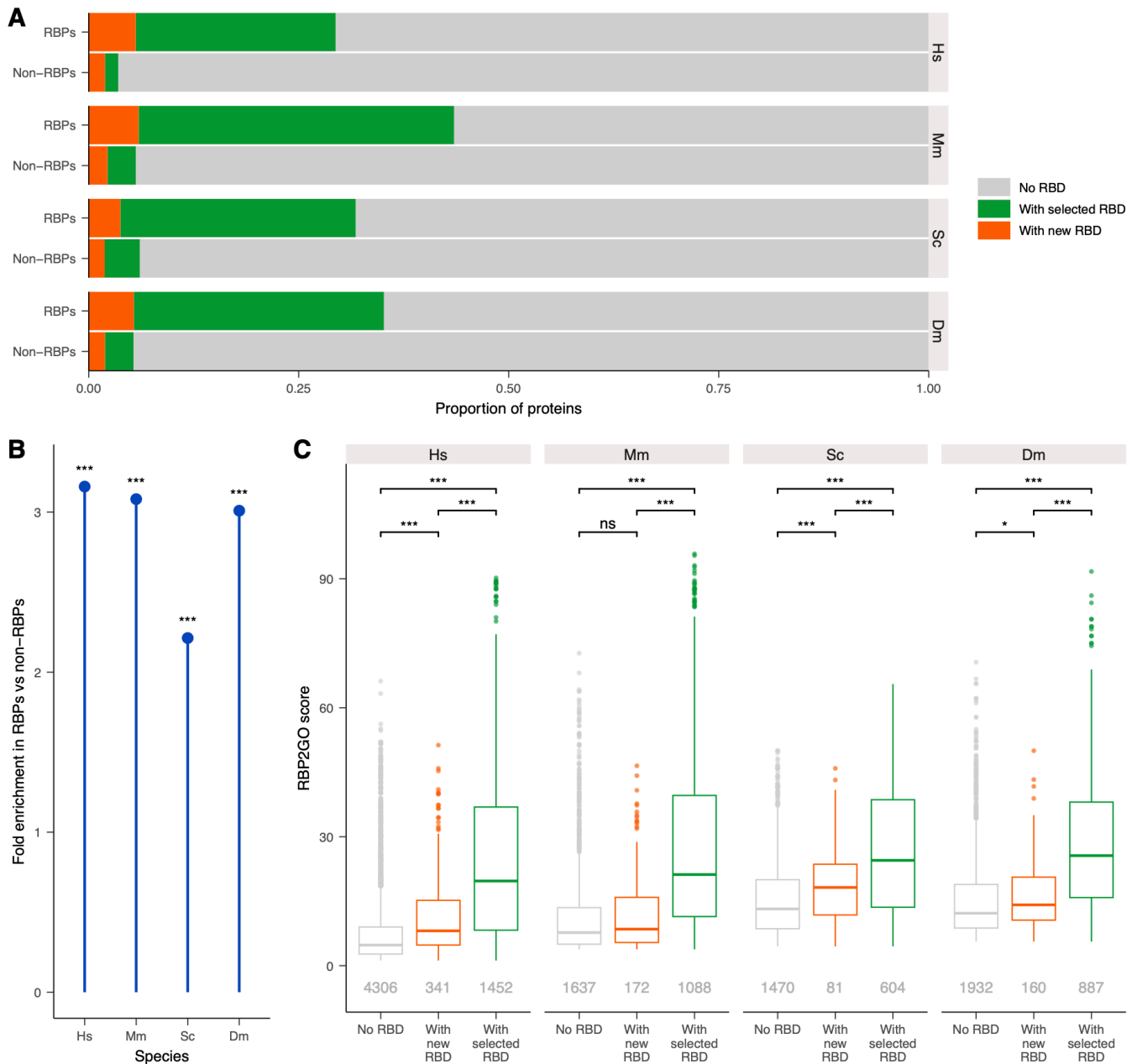


Figure 24: Validation of the new RBD candidates in the most studied species of the RBP2GO database

A. Proportion of proteins containing a known RBD (green), a newly predicted RBD candidate (orange) or neither (grey) in RBPs and non-RBPs. B. Enrichment of the newly identified RBD candidates in RBPs versus non-RBPs. *, **, *** and **** correspond to p-values < 0.05, 0.01, 0.001 and 0.0001, respectively, resulting from a Fisher's exact test. C. Boxplot showing the distribution of the RBP2GO score for the proteins containing an initially selected RBD, a newly discovered RBD or neither. *, **, *** and **** correspond to p-values < 0.05, 0.01, 0.001 and 0.0001, respectively, resulting from a Wilcoxon rank sum test.

2.9. Validation of the new RBDs using published data on RNA-binding peptides

Some of the studies compiled in the RBP2GO database used a technique allowing the identification of RNA-binding regions within RBPs (75, 76, 83, 123). These studies provide a list of the RNA-binding peptides they detect, as well as the proteins they belong to and the coordinates of the peptides within these proteins. I took advantage of this experimental data to validate the new RBDs.

In the workflow, the new RBDs were taken into account, along with the selected (Figure 5) and non-selected RBDs (present in the original list but which did not fulfill the selection criteria, Figure 5) which are present in more than five RBPs, to match the selection criteria of the new RBPs. Only the proteins containing one of these three types of RBD and could thus be unequivocally categorized into the groups on selected RBD, non-selected RBD or new RBD were used in this analysis. Then, the number of RNA-binding peptides from the aforementioned RBP screening studies overlapping at least to 50% with a domain was calculated for each different domain in each protein. This number was normalized by the total number of peptides overlapping in the protein, as well as by the coverage of the domain in the protein. Then, these normalized proportions were summed for a given domain across all proteins, and normalized by the total number of domains present in proteins containing at least one RNA-binding peptide (Figure 25A). The comparison of this number for all three groups of domains showed a significantly higher mean proportion of overlapping peptides in the new RBDs than in the non-selected RBDs, while it is lower than for selected RBDs, although this difference was not significant (Figure 25B). Only one of the 15 new RBDs did not overlap with any RNA-binding peptide, while it was the case for 17 selected and 10 non-selected RBDs. An example of a new domain presenting several overlapping RNA-binding peptides is the armadillo repeat, present in three proteins in our validation. These proteins contain 11 Armadillo repeats, among which six harbor a total of nine overlapping RNA-binding peptides (Figure 25C).

Overall, the published experimental data on RBDs confirmed that the 15 domains detected in the previous enrichment analysis are indeed RBDs, and they were integrated to the list of RBDs uploaded on the RBP2GO database, compiling a total of 992 InterPro IDs for RBDs.

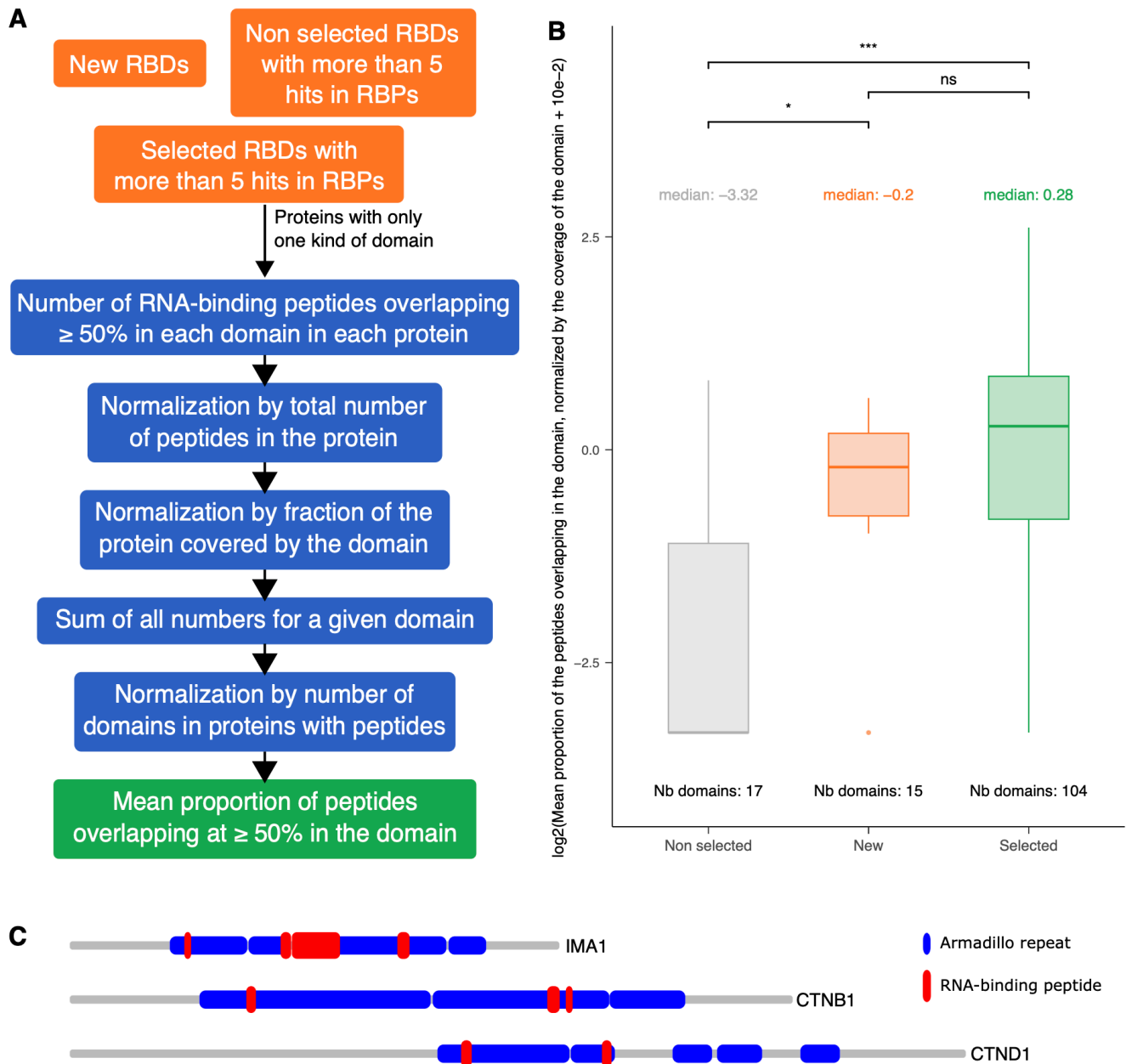


Figure 25: Validation of the new RBD candidates using experimentally identified RNA-binding peptides

A. Flowchart illustrating the steps to calculate the mean proportion of peptides overlapping in a given domain, depicted in B. B. Boxplot showing the distribution of the \log_{10} of the normalized mean proportion of the RNA-binding peptides overlapping at least to 50% with the domain + 1×10^{-7} , for each domain of the following categories: non-selected RBDs (grey), newly discovered RBDs (orange) and previously selected RBDs (green). The number above each box represents the median and the number under each box represents the total number of domains in this group. *, **, *** and **** correspond to p-values < 0.05, 0.01, 0.001 and 0.0001, respectively, resulting from a Wilcoxon rank sum test. C. Schematic representations of three proteins (grey) with Armadillo repeats (blue) and the overlapping RNA-binding peptides (red).

2.10. Establishment of a new RBP2GO composite score

In the previous sections, the presence of RBDs and Rfam IDs in a protein was shown to correlate well with its RNA-binding behavior, while the presence of disordered regions proved less valuable. To take these results into consideration and facilitate their use by the users of the RBP2GO database, a new score was created called the RBP2GO composite score. This score integrates the data already present in the RBP2GO score, but also the information given by the presence of RBDs and Rfam IDs (Figure 26A). The first half of the composite score is defined by the ratio between the number of times the protein was detected as an RBP and the number of datasets compiled in the database for the given species, normalized by 50. The second component of the new score is the mean of this ratio for the top 10 String interactors of the protein (102), normalized to 25. Finally, the third component of the composite score reflects the presence of RBDs and Rfam IDs. Since each of these annotations does not show the same enrichment in RBPs versus non-RBPs, they are attributed a quality factor to reflect this enrichment (Table 3). This quality factor ranges from 1 to 5, and is higher when the enrichment of the InterPro ID in RBPs versus non-RBPs is higher. For the InterPro IDs whose enrichment is infinite, the quality factor is attributed depending on the number of RBPs annotated with this ID. Finally, the different quality factors are added for all RBDs and Rfam IDs present in the protein, and the sum is limited to 25, resulting in the third component of the composite score (Figure 26A).

Ratio RBP/non-RBP	Number of hits in RBPs	Quality factor of the RBD or Rfam ID
< 2	Not taken into account	1
≥ 2		2
≥ 4		3
≥ 8		4
≥ 16		5
Inf	≤ 2	2
Inf	≤ 4	3
Inf	≤ 8	4
Inf	> 8	5

Table 3: Attribution rules of quality factors to RBDs and Rfam IDs

Ratio RBP/non-RBP = ratio of the number of hits in RBPs and the number of hits in non-RBPs for a given InterPro ID; Inf = infinite since InterPro ID not detected in non-RBPs

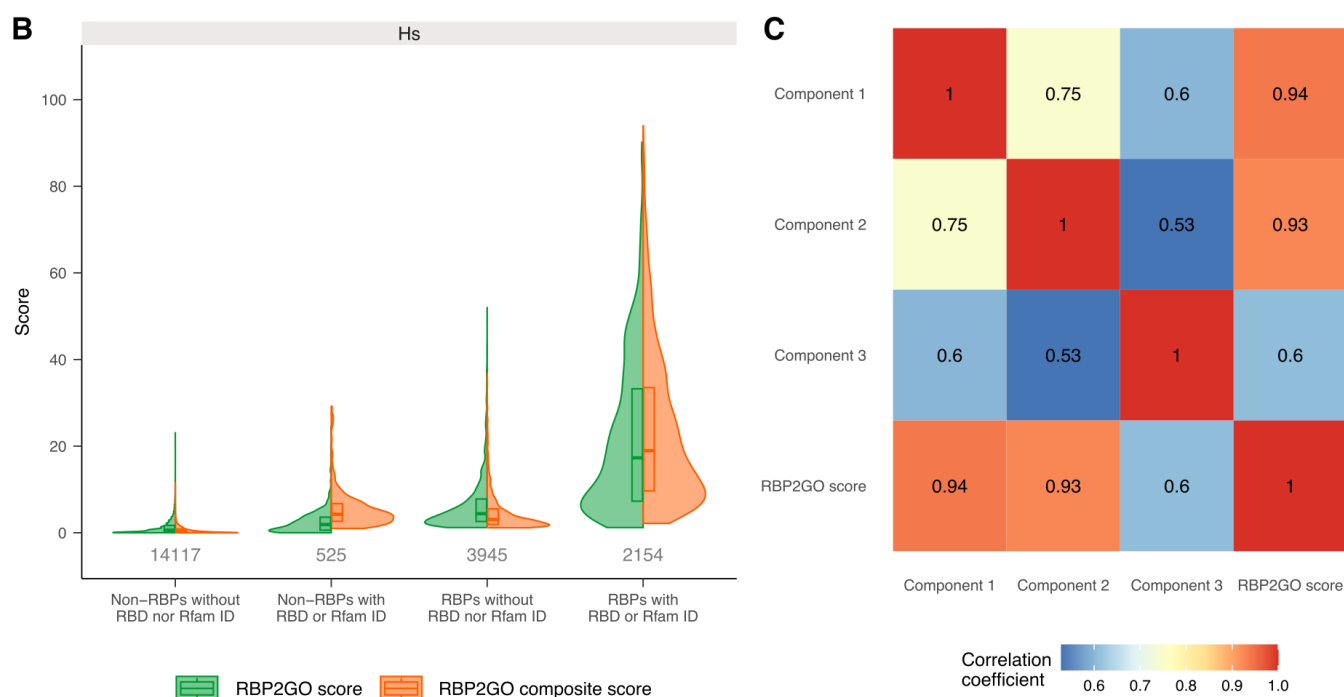
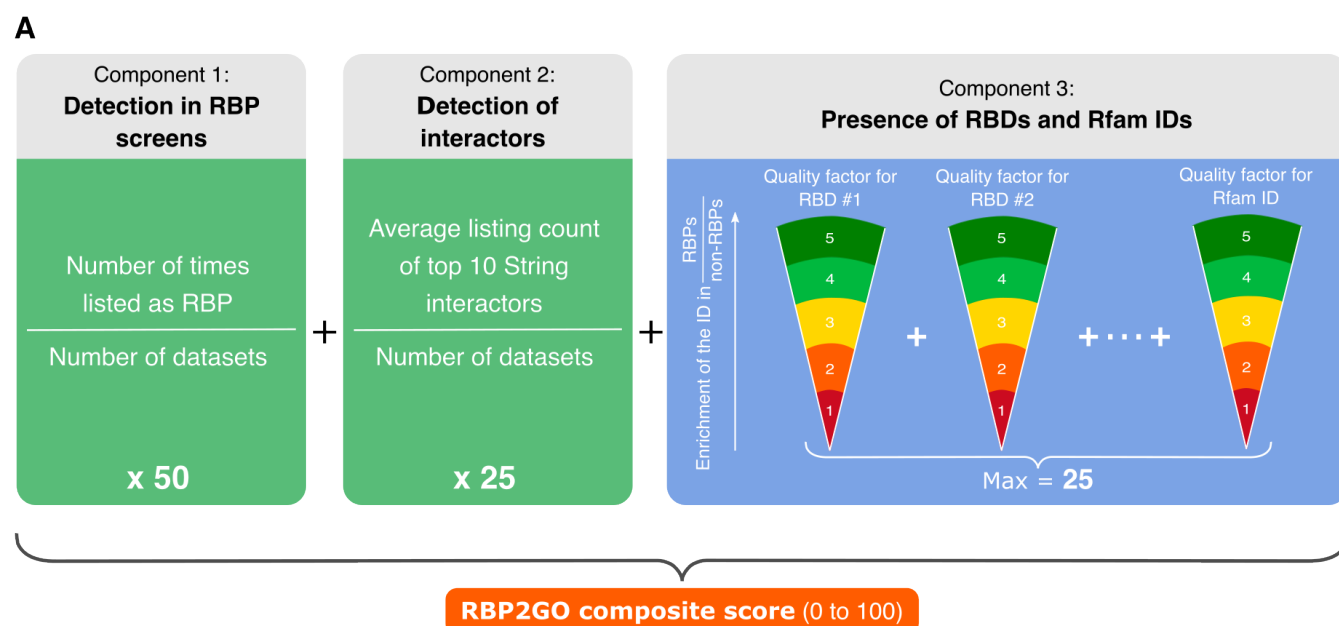


Figure 26: Development of the new RBP2GO composite score

A. Schematic representation of the calculation of the RBP2GO composite score. The attribution rules of the quality factors can be found in Table 3. B. Violin plot representing the distribution of the RBP2GO score (green) and the new RBP2GO composite score (orange) for non-RBPs without RBD, non-RBPs with RBD, RBPs without RBD and RBPs with RBD. C. Pearson's correlation factors between the three components of the new RBP2GO composite score and the RBP2GO score. The lowest scores are depicted in blue and the highest scores in red.

The RBP2GO composite score better separated the non-RBPs and RBPs with no RBD or Rfam ID from the non-RBPs and the RBPs with at least one annotation type, respectively (Figure 26B). Furthermore, the three components of the composite score showed a correlation coefficient between 0.53 and 0.75, showing they correlate together but are not redundant. This shows that the new component 3 especially, based on RBDs and Rfam IDs, added independent information (Figure 26C). This new RBP2GO composite score thus represents an interesting tool to judge the RNA-binding abilities of a given protein in the RBP2GO database based on experimental data as well as domain composition and protein families.

2.11. Implementation of the new data in the RBP2GO database

Finally, I integrated all the data generated in this thesis into the RBP2GO database. The homepage was modified to display information on the RBDs, the Rfam IDs and the RBD-containing non-RBP proteins (here called RBP aspirants, Figure 27A). New search options were also added to the advanced search in all species (Figure 27B). The users now have the possibility to search for proteins containing a list of specific InterPro IDs, and to limit their search to proteins with RBDs and/or with Rfam IDs. The new RBP2GO composite score was also added to the advanced search options, so users can refine the results of the search based on a specific score. When searching for a protein using the “Protein Search” tab, they are now classified according to the RBP2GO composite score (Figure 27C).

A new “Domain Information” subtab was created for each individual protein, and displays detailed information about the RNA-binding features of the selected protein. The presence of RBDs and Rfam IDs, as well as disordered regions and the disordered fraction are visible, and links to the InterPro (145) and the MobiDB (143) databases are provided. A first table displays the information on the InterPro annotations taken into account in this study (Domain, Repeat and Family types of annotations), with their RNA-binding behavior. The number of domains and repeats present in the selected protein are also provided. The other InterPro IDs present in the protein are present in a second table below the first one (Figure 28). Both tables can be downloaded by clicking on the “Download Table” button.

Moreover, to provide more information on the RBDs and Rfam IDs and how they were selected, a new section called “Search domains” was created in the sidebar (Figure 29A). This section contains three different tabs. The first two tabs provide information on the selected RBDs and Rfam IDs, respectively, as well as their selection and their distribution in the proteins of the database. The last one, “Domain search”, allows the user to search for a given InterPro ID and/or a list of InterPro IDs and see if they are RNA-binding or not (i.e., if they are present in our list of selected RBDs and Rfam IDs). The lists of selected RBDs (including the newly identified ones) and Rfam IDs were added to the “Download” section for an easy access (Figure 29B). A list of high-confidence human RBPs is also provided in this section based on the RBP2GO composite score.


A

Welcome to RBP2GO

RBP2GO is a comprehensive database dedicated to the analysis of RNA-binding proteins, their interactions and functions. RBP2GO offers three search options for each species, which can be accessed via the sidebar under **SEARCH SPECIES**. Information on RNA-binding domains (RBDs) and RNA-related family IDs (Rfam IDs) is also available under **SEARCH DOMAINS**. More details can be found in the **HELP** menu.


RBP candidates
22554

RBP aspirants
7729




RNA-binding domains
992

RNA-related family IDs (Rfam IDs)
672



Species
13



B

Protein Search **GO Search** Advanced Search

Advanced search option including search for list of proteins or GO terms.

Input Format

Protein List
Paste list of proteins here. One per line!

Limited to CGC Genes

Limited to RBD-containing proteins

Limited to proteins with RNA-related family IDs (Rfam IDs)

GO List
Paste list of GO terms here. One per line!

InterPro ID List
Paste list of InterPro IDs here. One per line!

OR

OR

RBP2GO Composite Score
0 100

RBP2GO Score
0 100

Isoelectric Point (pI)
0 14

Submit here! **Reset**

C

Search Result

There are several matches for: 'hnrp*'

There are: 29 **RBP2GO proteins** and 1 **other proteins**.

RBP2GO Proteins	Other Proteins	
Entry Name	Gene Name	RBP2GO Composite Score
HNRPL_HUMAN	HNRNPL	94
HNRPK_HUMAN	HNRNPK	93
HNRPM_HUMAN	HNRNPM	86.6
HNRH1_HUMAN	HNRNPH1	86
ROA1_HUMAN	HNRNPA1	85.7
HNRPR_HUMAN	HNRNPR	85.3
HNRPF_HUMAN	HNRNPF	84.5
HNRH2_HUMAN	HNRNPH2	83.6
ROA3_HUMAN	HNRNPA3	81.2
ROA2_HUMAN	HNRNPA2B1	80.6
ROA0_HUMAN	HNRNPA0	78.8
HNRH3_HUMAN	HNRNPH3	78.5
HNRPD_HUMAN	HNRNPD	65.8
HNRDL_HUMAN	HNRNPDL	64.6
HNRPC_HUMAN	HNRNPC	60.4

Figure 27: Update of the RBP2GO database to integrate the new data

A. Screenshot of the new homepage of the updated RBP2GO database. B. Screenshot of the new “Advanced search” panel. C. Example of search results from the “Protein Search” tab, using the terms “hnrp*”.

[Heterogeneous nuclear ribonucleoprotein L](#)

Number of non-overlapping RBDs:	4	
RBDs content fraction:	0.58	
RNA-related family ID (Rfam ID):	Yes	
Disordered region:	Yes	Link to mobiDB
Disordered fraction:	0.33	

Domains, repeats and family-IDs

Show **5** entries

Search:

Name	ID	Type	RNA-binding	RBD number
HnRNP-L/PTB	IPR006536	Family	Yes	
hnRNP-L, RNA recognition motif 3	IPR034816	Domain	Yes	1
hnRNP-L, RNA recognition motif 1	IPR035005	Domain	Yes	1
hnRNP-L, RNA recognition motif 2	IPR035008	Domain	Yes	1
hnRNP-L, RNA recognition motif 4	IPR034817	Domain	Yes	1

Showing 1 to 5 of 7 entries

Previous **1** 2 Next

[Download Table](#)

Other InterPro annotations

Show **5** entries

Search:

Name	ID	Type
Nucleotide-binding alpha-beta plait domain superfamily	IPR012677	Homologous_superfamily
RNA-binding domain superfamily	IPR035979	Homologous_superfamily

Showing 1 to 2 of 2 entries

Previous **1** Next

Figure 28: Information about the RBDs and the Rfam IDs integrated in the RBP2GO database. Screenshot of the new “Domain Information tab”.

A

RBD Information Rfam ID Information **Domain Search**

In this panel, it is possible to search for **protein domains** in order to **determine** whether or not they are **RBDs**.

InterPro ID List

IPR000504
IPR001962

[Submit here!](#) [Reset](#)

The **RNA-binding status** is only available for the Domain, Repeat and Family types of InterPro IDs.

If it is intended to search for proteins with specific domains, please use the **Advanced Search** in the respective species and use the **InterPro ID List** field.

Show **10** entries Search:

ID	Type	Name	RNA-binding
All	All	All	All
IPR000504	Domain	RNA recognition motif domain	Yes
IPR001962	Domain	Asparagine synthase	No

Showing 1 to 2 of 2 entries Previous **1** Next

[Download Table](#)

B

 Domains [List of selected RBDs](#) [List of selected Rfam IDs](#) [List of high-confidence RBPs](#)

 R scripts [Script of the database](#) [Script of the analysis](#)

Figure 29: Integration of the generated data into the RBP2GO database

A. Screenshot of the new homepage of the updated RBP2GO database. B. Screenshot of the new “Advanced search” panel.

In a nutshell, the RBP2GO database was adapted to integrate all the information about RNA-binding domains that were gathered in this study, and now provides a view on experimental RBP screens as well as the RNA-binding domains and family ties of each protein.

3. Discussion

In this study, I exploited the compilation of studies available in the RBP2GO database (108) to analyze the presence of RNA-related InterPro annotations, as well as disordered regions, in the proteins that have been experimentally detected as RNA-binding. The gained knowledge was then used to find new RBP and RBD candidates, expanding the available data on RBPs without performing supplementary experiments. A new RBP2GO composite score was also created to have one indicator merging all the experimental data, domain content and protein family information available for each protein. The RBP2GO database was also updated to integrate the newly acquired information in a user-friendly way.

To have the most comprehensive list of RNA-related InterPro annotations, both published lists of RBDs and lists downloaded from the InterPro website were used. Three studies were compiled, using data both from a literature-mining study (148) and experimental results (68, 75). Regarding the information gathered from InterPro, three different strategies were used to compile the maximum of InterPro IDs that have been annotated as RNA-binding, naming a keyword search, a search on the GO terms and the domains overlapping with the “RNA binding domain homologous superfamily” ID. The total of 2712 InterPro IDs was first filtered to keep the annotations related to specific domains, namely “Domain” and “Repeat” InterPro IDs. The aim was then to select the domains that were enriched in the RBPs of the database compared to the non-RBPs. To this end, the selection was split in three parts to select in all the species of the database, in decreasing order depending on the number of datasets available. This provided a list of 977 RBDs, selected based on experimental datasets. This list is then less biased than the lists of RBDs that have been published so far, since it does not rely on a single experimental setup and combines different types of experiments. It is also more up-to-date than the list of RBDs from Gerstberger et al. (148), which is widely used to validate experimental results, but is now almost a decade old.

However, some domains that have recently been proven to bind RNA were not selected in this study. This is the case e.g. for the HMGB (high mobility group box) domain, a DNA-binding domain for which recent studies proved that is also binds to RNA (156, 157). This domain was present in the initial list of InterPro IDs. Nonetheless, since it was present in 28 human non-RBPs and 27 human RBPs, it was not selected as an RBD, but very close to the cut-off. This shows that the selection process is quite stringent, and more oriented towards the isolation of

domains that bind to RNA as their primary function. The list of selected RBDs proposed here can thus be considered as a reliable list of domains that possess a high binding affinity for RNA, and could be, for example, used to validate results of a proteome-wide RBP screen, without validating false-positives. The use of a unified list in the field could also help to better compare different experimental strategies by providing a common base for validation.

The previous point also suggests that different lists of RBDs, with different affinities, could be created, by running the selection process with different frequency ratio thresholds. It would then be interesting to evaluate the distribution of these different classes of RBDs in the RBPs, and maybe help to better understand why two-thirds of the RBPs in the most studied species did not exhibit any RBD in the present analysis.

Although 852 out of the 977 selected RBDs were selected in Hs in the first selection step, the proteins with an RBD displayed a higher RBP2GO score than their counterparts with no RBD in all of the studied species. This shows that the knowledge acquired in the most studied species can be transferred and used in other species. Moreover, the RBP2GO score correlated with the number of RBDs present in RBPs, and RBPs showed a higher RBD content fraction than non-RBPs. This is in accordance with the knowledge that single RBDs are rarely specific, and thus several of them are cooperating in the same proteins to bind RNA (109).

Nevertheless, all of the species contained at least 200 RBDs, and the species with the least number of domains were also the most evolutionary remote organisms from Hs, like Ec, Tb, ST, Lm and Ld. More than half of the datasets compiled on RBP2GO concern experiments done in human or mouse cells, showing a clear lack of knowledge on prokaryotes and unicellular organisms. This was also highlighted by the higher proportion of non-RBPs containing an RBD in the least studied species.

One notable exception to this statement is Ec, which had only 48 non-RBPs with no RBD, compared to Dr that had 1729, and the same number of datasets. This can be explained by one study using the TRAPP technology, which detected more than 91% of the RBPs in Ec, and more than 71.3% of the RBD-containing proteins. However, it also detected 91.3% of the RBPs with no RBD, therefore this technique proved very sensitive but less specific than the ones used for the other experimental datasets.

To fully exploit the information provided by the InterPro annotations, the 1028 family IDs of the initial list were also studied. These annotations provide information on the family to which a protein belongs, but do not give details on a specific domain. The same selection process was used as for the RBDs, resulting in 672 Rfam IDs. Since the same selection criteria were applied,

this list of RNA-binding protein families can be considered as quite stringent, as it contains families that were obligatory more abundant in RBPs than in non-RBPs. It could then be considered as well as a reliable list to perform validation of experimental studies, and used as an easy-to-access consensus.

The Rfam IDs proved to be interesting in the evaluation of RBP candidates, as their presence also correlates with a higher RBP2GO score. Here as well, most of the Rfam IDs were selected in humans, but the correlation with the score in all species showed the possibility to transfer this knowledge to other organisms. However, the proportion of RBPs containing only Rfam IDs and no RBD was quite low, and did not cover a significant amount of the RBPs with no RBD.

The lack of knowledge in poorly studied species highlighted by the RBDs was also underscored by the Rfam IDs. Again, more non-RBPs with Rfam IDs were present in these species, and their score was higher than the non-RBPs with no Rfam IDs. This means that their interactions partners have a higher RBP2GO score, and is in agreement with the previously published tendency of RBPs to interact with other RBPs (103).

The last type of domains that was scrutinized in this study are intrinsically disordered regions. To be able to calculate disordered content fractions, I needed to access disorder predictions with given coordinates for the different IDRs. Furthermore, the predictions needed to be rather specific and conservative, and ideally provide several different types of disorder, as some specific types were shown to bind to RNA (127). As a result, the predictions from the MobiDB-lite algorithm were used for this analysis (142, 143, 147).

Almost no data on IDRs was available for *Ec* and *ST*, so these species were not included in the analysis regarding disorder. There are some studies that suggest that the proportion of intrinsic disorder increased sharply during the transition from prokaryotes to eukaryotes (158, 159). This could explain the lack of disorder predictions available for *Ec* and *ST*.

Surprisingly, the RBD-containing non-RBPs and RBPs showed a higher content in disordered regions than their counterparts without RBDs. Since some proteins are able to bind RNA only via their IDR, such as NKAP (127, 139), it was expected that the RBPs with no RBD would have a higher disordered content, which would also have explained their lack of known RBDs. However, several well-known RBPs are known to contain both RBDs and IDRs, such as Fus (75). In this protein, both the RRM and the IDR of the prion-like domain are necessary for specific RNA-binding (138). Moreover, IDRs can also serve as linkers between known RBDs, conferring them enough flexibility to arrange themselves on the RNA target. This enhances the

size of the recognized sequence but also allows the protein to recognize a specific 3D structure (109). The results of this analysis thus support a widespread cooperation between IDRs and RBDs in the binding of RNA by a protein.

However, in human and mouse, RBPs with no RBD and a high RBP2GO score showed an enrichment in IDRs compared to RBPs with no RBD and a low RBP2GO score. This indicates that, in these two species only, some RBPs with no known RBD might bind RNA through disordered regions alone. But this mode of binding seems to be more an exception, and can be considered for proteins with a high RBP2GO score only.

Several subtypes of disorder are also given as an output from the MobiDB-lite algorithm. I first analyzed the presence of polyampholyte IDRs, namely IDRs with both positively and negatively charged amino acids, since it is a feature of RNA-binding IDRs. However, the results were the same as for the general IDR predictions, with an enrichment in RBD-containing RBPs compared to RBPs with no RBD (24, 127). The results for positive-polyelectrolyte and negative-polyelectrolyte IDRs were not shown but displayed the same enrichment. The only subtype of disorder which was enriched in RBPs with no RBD compared to the RBD-containing ones are coiled-coil domains, and only in Hs, Mm, Sc and Dm. These domains are involved in neuronal granule proteins, and provide the ability to phase-separate (154). However, no difference could be observed between the content in coiled-coil regions in the non-RBPs without an RBD and the RBPs without an RBD, which puts into question the involvement of these regions in RNA-binding in the proteins with no RBD.

Overall, the presence of IDRs did not correlate well with the RBP2GO score of the proteins of the RBP2GO database, and was thus not considered for the new RBP2GO composite score. The data concerning the presence of disorder were nevertheless integrated into the new version of the database, with a link redirecting to the MobiDB website for more information.

Since no enrichment in disorder could be found in the two thirds of the RBPs that do not contain any RBD, this raised the question of the presence of unknown RBDs in these proteins. An enrichment analysis was thus performed in the RBPs with no RBD and a high RBP2GO score compared to three reference datasets, and resulted in the identification of 15 new candidate RBDs. These domains showed an enrichment in Ms, Sc and Dm, as well, even though they were selected from human proteins only.

In addition, some of these domains have already been linked to RNA-binding or RBPs in the literature. Concerning the repeats, the WD40 repeats have been found in several RBPs, notably proteins involved in RNA-processing like TATA-box binding proteins and transcription-

associated factors (160). WDR43 (WD40-repeats containing protein 43), a protein involved in the processing of rRNA and the ribosome biogenesis, also contains WD40 repeats (161). The tetratricopeptide repeat was also detected in the enrichment analysis. These repeats are present in the IFIT (Interferon-induced with tetratricopeptide repeats) proteins, which are known to bind RNA, and do not exhibit any other domain on InterPro (162, 163).

The cyclophilin-type peptidyl-prolyl cis-trans isomerase domain is present in cyclophilins, a type of peptidyl-prolyl isomerase. Some of these proteins are present in the nucleus and part of spliceosomal complexes, where they can interact with non-coding RNAs involved in splicing (164). For example, Rct1 is involved in the processing of pericentromeric transcripts into siRNAs (165).

The Armadillo (Arm) and atypical Armadillo repeats were also identified as new RBDs and can be found in importins. The importin-alpha, importin-beta binding domain was also part of the 15 new RBDs. Both importins are involved in the regulation of RNA export in the cytoplasm (166), and other Armadillo-repeat containing (ARMC) proteins are important for RNA localization and mRNA regulation (167).

Finally, the enrichment analysis also highlighted the Histidine kinase/HSP90-like ATPase domain as a new RBD. HSP proteins are molecular chaperones, ensuring the correct folding of other proteins (121). But they were also recently identified as RNA-binding proteins. Hsp90 plays a role in the loading of RNAs into the RISC complex (124), while Hsp70 can bind AU-rich elements in mRNAs and influence their translation (125, 126).

To further validate these 15 RBDs, I used the studies already compiled in the RBP2GO database that provide information on RNA-binding peptides (75, 76, 83, 123). I calculated a mean proportion of overlapping peptides for the new RBDs, as well as selected and non-selected RBDs. The new RBDs showed a higher mean of overlapping peptides than the unselected RBDs, validating them with experimental data. Furthermore, the data was generated with different experimental approaches, limiting potential biases. An additional validation that could be done in the future would be to perform a CLIP experiment on proteins containing these domains and no other RBDs, or even on the fraction of the protein containing the domain of interest, to confirm a direct interaction with RNA.

However, these new RBDs are not present in a large proportion of the RBPs with no RBD. These other proteins from this group thus remain with no known RNA-binding feature or RNA-related annotation.

There are several hypotheses that could explain this large proportion of RBPs that do not exhibit RNA-binding features.

The first explanation could be the stringency of the RBD selection. As seen before, some domains known to bind to RNA, such as the HMGB domain, were not selected as RBDs for this analysis. These RBPs could then bind RNA but through domains for which RNA-binding is not their primary function. The domain could then be present in a significant number of proteins that do not bind RNA and not be selected. They also may be binding RNA through domains that are not yet known to bind to RNA, and that were then not integrated into the initial list of InterPro IDs. A recent study showed that a high proportion of the RBPs with no canonical RBD do not exhibit any sequence specificity. The authors thus propose to create subcategories to characterize RBPs: specific RBPs, unspecific RBPs and RNA-associated proteins (168). The RBPs with no RBD would then fall in one of the two latter categories.

Secondly, these proteins might not be binding RNA directly but rather be RNA-dependent proteins, so proteins that bind RBPs. In addition, two of the human datasets integrated in the RBP2GO database were generated using technologies detecting RNA-dependent proteins (89, 90). UV crosslinking techniques are generally admitted to only crosslink RNA and proteins that are in contact with them, but the maximum crosslinking distance can vary between experimental setups, and the technique could allow proximal proteins that do not bind the RNA directly to still be covalently bound to it (169, 170). These proteins should have a lower listing count (number of times it was detected as RBP) and a higher mean listing count for their interactors, and could therefore be isolated from the others by comparing these two numbers.

Finally, some proteins could be binding unspecifically to the material used for the isolation of RNA-protein complexes, like magnetic beads. Indeed, these beads are prone to unspecific binding by diverse proteins and RNAs (171), and it would be interesting to have control experiments to evaluate the subsequent level of background.

Interestingly, 52% of the human RBPs with no RBD or Rfam IDs exhibit a listing count of one, which means they have been detected as RBPs in only one dataset. When looking at the number of these proteins for each of the 43 human datasets, 72.7% were detected with PTex (85), R-DeeP (89), or SONAR (103) (Figure 30A). In the case of SONAR, the RBPs were predicted using their interactors, based on the principle that the RBPs tend to interact with other RBPs (103). Thus, it is most susceptible to detected RNA-dependent proteins, that do not interact directly with RNA. R-DeeP was also developed specifically to detect such proteins. However, most of the RBPs with no RNA-binding feature detected in these studies exhibit a low average for the listing count of their top 10 String interactors (Figure 30B), which is not in accordance with these proteins being RNA-dependent proteins. Hence, the techniques detecting RNA-dependent proteins seem more likely to produce false positives. The RBPs detected only by

PTex also exhibit a low average listing count for their interactors (Figure 30B). However, two other techniques based on the same experimental principle as Ptex, called OOPS (83) and XRNAX (84), do not show as many proteins with no RNA-binding feature and a listing count of one. Thus, PTex generates more false positives than other protocols based on the same principle. This shows that proteins with a very low RBP2GO composite score, even though they have been detected as RBPs, should be considered false positives. Moreover, it does not appear that one experimental principle exhibits a better specificity than the others for the reliable detection of RBPs, but some studies definitely generate more background than others.

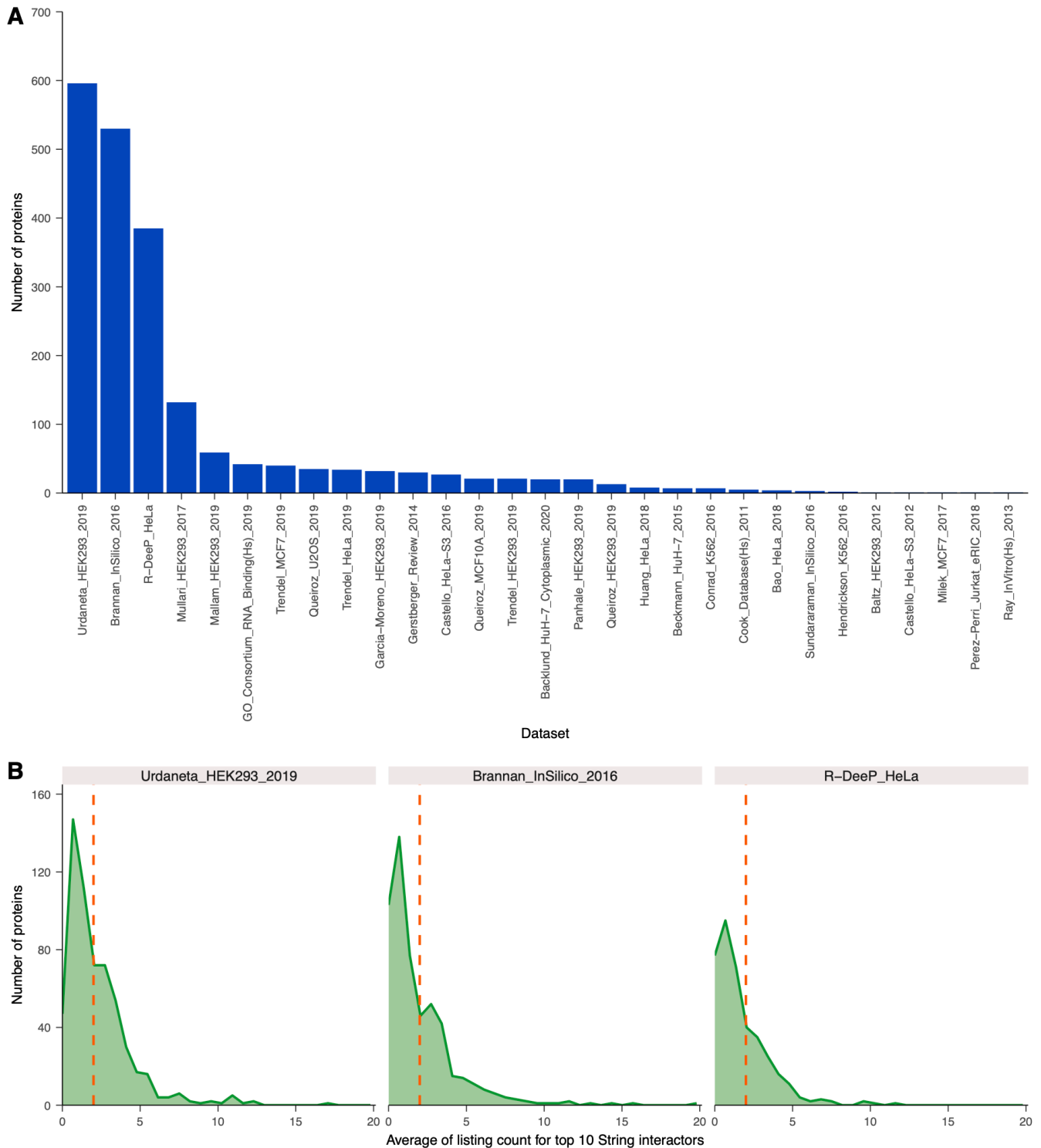


Figure 30: Repartition of human RBPs with no RBD and detected only once in the human datasets

A. Bar graph showing the number of RBPs that were detected only once (listing count or component one of the RBP2GO composite score of 1) and containing no RBD for each dataset of the RBP2GO database. The dataset containing no such protein are not represented. B. Density plot showing the distribution of the RBPs with no RBD and a listing count of one depending on the average listing count for the top 10 String interactors of each protein (component two of the RBP2GO composite score) for three human datasets. The red line represents an average of 2.

To help future users of the database better discriminate between all the listed RBP candidates by having one metric grouping experimental, interaction, domain and family data, a new RBP2GO composite score was created. Half of this score is based on the detection of proteins in the proteome-wide RBP screening datasets compiled in the database, and half of it on the orthogonal data, the detection as RBP of the interactors and the presence of RBDs and/or Rfam IDs. Since the presence of disorder was not clearly a decisive feature of RBPs, it was not integrated into the score. Using this score, a list of high-confidence human RBPs was selected. Their RBP2GO composite score is higher than 10, and a maximum of one of the three components of the score can be null. This list of 2019 proteins can be used in the future as a consensus, and is available for download on the RBP2GO website.

Finally, all of the information gathered on the proteins of the database were integrated in a new version of RBP2GO. New dedicated sections were created to provide an easy access to this information to the rest of the scientific community. All data tables can be downloaded, and information on how the data was selected is provided. In the future, the newly published proteome-wide screens will be gathered and added to the database. All of the information should be updated to the newest released versions, and the RBP2GO scores should be calculated again, to provide up-to-date data on RBPs.

4. Materials

All of the analyses and the figures, as well as the RBP2GO database, were performed with R, version 4.1.2 (2021-11-01) "Bird Hippie".

Package	Version
janitor	2.2.0
pals	1.7
curl	5.0.0
rbioapi	0.7.7
httr	1.4.4
gridExtra	2.3
ggbreak	0.1.1
ggsignif	0.6.4
ggpubr	0.6.0
valr	0.6.6
jsonlite	1.8.4
readxl	1.4.2
forcats	1.0.0
stringr	1.5.0
dplyr	1.1.0
purrr	1.0.1
readr	2.1.4
tidyr	1.3.0
tibble	3.1.8
ggplot2	3.4.1
tidyverse	1.3.2

Table 4: List of the R packages used for the analysis presented in this thesis

Package	Version
ggupset	0.3.0
rstatix	0.7.2
scales	1.2.1
gridExtra	2.3
ggrepel	0.9.3
ggbreak	0.1.1
ggsignif	0.6.4
ggpubr	0.6.0
forcats	1.0.0
stringr	1.5.0
dplyr	1.1.0
purrr	1.0.1
readr	2.1.4
tidyr	1.3.0
tibble	3.1.8
ggplot2	3.4.1
tidyverse	1.3.2

Table 5: List of the R packages used to generate the figures presented in this thesis

Package	Version
dplyr	1.1.0
tidyr	1.3.0
shinybusy	0.3.1
formattable	0.2.1
shinycssloaders	1.0.0
rmarkdown	2.2
magrittr	2.0.3
DT	0.27
lattice	0.20-45
gridExtra	2.3
ggplot2	3.4.1
raster	3.6-14
sp	1.6-0
shinyWidgets	0.7.6
shinydashboard	0.7.2
rintrojs	0.3.2
shinyBS	0.61.1
shinyjs	2.1.0
shiny	1.7.4

Table 6: List of the R packages used for the RBP2GO database

5. Methods

5.1. Compilation of a list of RNA-binding domain candidates

The lists of RBDs were downloaded from the respective supplementary information of three studies published between 2012 and 2016 (68, 75, 148). One of them is comprised of manually curated domains (148), while the two others come from the results of proteome-wide RBP screens (68, 75). Regarding the study from 2016, only the domains with an adjusted p-value lower than 0.05, hence the significant domains, were selected. The domain IDs were converted to InterPro IDs if referenced from another database such as Pfam (35) and compiled with the list of IDs from InterPro. The outdated IDs were updated or discarded if no corresponding InterPro ID could be found. Altogether, 808 IDs were retrieved from the published datasets. A keyword search on the InterPro website with the words “RNA binding” and “RNA-binding” showed that the last option retrieved the most results. Hence, the results of the keyword search with the terms “RNA-binding” was filtered to keep only InterPro IDs and was downloaded from the InterPro website (<https://www.ebi.ac.uk/interpro/>, InterPro v.88 released on the 10th March 2022) (32). Since 2012, InterPro IDs have been manually annotated with Gene Ontology (GO) terms (149), thus, I also downloaded the list of IDs annotated with the GO term “RNA-binding” (GO:0003723). Finally, the “RNA binding domain superfamily” (IPR035979) overlaps with several other domains. Accordingly, I downloaded the list of these overlapping domains from the InterPro website. Taken together, these three lists amounted to 2251 unique InterPro IDs.

The IDs from InterPro and the published datasets were combined together in a list of RNA-related IDs, amounting to 2712 unique InterPro IDs. InterPro provides different entry types: “Domain”, “Family”, “Homologous Superfamily”, “Repeat”, “Site” and “Unintegrated” (145). Since the focus of this section is the RNA-binding domains, the list of InterPro IDs was filtered to keep only the “Domain” and “Repeat” types of IDs. This resulted in a list of 1289 RBD candidates, that were subsequently submitted to a selection process.

5.2. Selection of the RNA-binding domains

The aim of this selection was to keep the InterPro IDs that are enriched in RBPs compared to non-RBPs. Furthermore, all the species of the RBP2GO database do not exhibit the same

coverage in terms of number of datasets (Figure 6), so the selection process was designed to take this information into consideration.

Hence, a three-step selection procedure was applied to the set of RBD candidates based on the hit ratios between RBPs (RBP candidates from proteome-wide studies) and non-RBPs in the species reported in the RBP2GO database (<https://rbp2go.dkfz.de>) (20). The InterPro IDs from the list of RBD candidates that were enriched (hits in RBPs > hits in non-RBPs, i.e. more often found in RBPs than in non-RBPs) in *Homo sapiens* (Hs) were first selected (Selection 1, Figure 5A), as the number of studies available for this species is by far the largest (Figure 6) as compared to the other species. If required, the UniProt IDs of the proteins were updated (UniProt 2022_01 release from the 23rd February 2022) (144). However, if two Uniprot IDs corresponding to RBPs have been fused in this version, both were kept in the dataset, since the deletion of an RBP would result in a loss of listing count data.

In a second step (Selection 2, Figure 5A), the ratio between the number of hits in RBPs and the number of hits in non-RBPs for the RBD candidates that were not present in Hs was calculated for *Mus musculus* (Mm), *Saccharomyces cerevisiae* (Sc) and *Drosophila melanogaster* (Dm). The InterPro IDs with a ratio higher than one in at least half of the species in which they could be found were selected. Finally (Selection 3, Figure 5A), within the RBD candidates that were not present in Hs, Mm, Sc and Dm, the RBDs that were enriched in at least half of the remaining nine species in which they were present, i.e., *Arabidopsis thaliana* (At), *Caenorhabditis elegans* (Ce), *Plasmodium falciparum* (Pf), *Escherichia coli* (Ec), *Danio rerio* (Dr), *Trypanosoma brucei* (Tb), *Salmonella Typhimurium* (ST), *Leishmania mexicana* (Lm) and *Leishmania donovani* (Ld) were selected.

This three-step selection procedure led to the selection of 977 InterPro IDs, that will be henceforth called “RBDs” or “selected RBDs”. The proteins were then analyzed for the presence of these RBDs, and the results were displayed in Figures 7 to 10.

5.3. Protein expression levels in HeLa cells

To better understand the correlation between the RBP2GO score and the expression of the proteins in humans, expression data from a deep proteome analysis were downloaded from the EBI atlas website (<https://www.ebi.ac.uk/gxa/experiments/E-PROT-19/Results>, (150)). This

experiment was performed using mass spectrometry-based shotgun proteomics, and provides expression data for each protein in parts per billion. This data was displayed in the Figures 8B and 8C, and the proteins that had no expression data were not taken into account.

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

The InterPro database provides the coordinates of each InterPro ID for each Uniprot ID available on their website. The file containing these coordinates was downloaded from the InterPro database (<https://ftp.ebi.ac.uk/pub/databases/interpro/releases/88.0/>, released on the 10th March 2022, (145)), and then reduced in size in the command terminal using the “rg” (ripgrep) function to only keep information for the proteins of the RBP2GO database and allow to open it in R. Then, as several coordinates are available for each InterPro ID in each protein, I used the “bedmerge” function of the “valr” package (172) to merge all the overlapping annotations and thus keep only one set of coordinates for each InterPro annotation.

To compute the fraction of the protein length covered by RBDs, or RBD content fraction, I merged separately all the overlapping RBDs present in a protein. Next, I used the protein length previously retrieved from UniProt (https://ftp.ebi.ac.uk/pub/databases/uniprot/previous_releases/release-2022_01/, released on the 23rd February 2022 (144)) to calculate the fraction of the sequence annotated as RBD for each protein.

To compute the number of RBDs, I merged the coordinates of the InterPro IDs overlapping by more than 10 amino acids using the “bedmerge” function of the “valr” package again for each protein. The number of coordinate pairs (start, end) generated was then counted to obtain the number of RBDs per protein.

The results of these analyses were used to generate the Figure 11.

5.5. Construction of a list of RNA-related family IDs and selection of the IDs enriched in RBPs.

As was said previously, the InterPro database also provides annotations related to the protein families in the form of “Family” InterPro IDs. Some of these annotations were present in the

lists of published RBDs, so I selected a list of RNA-related InterPro family IDs. To this aim, the list of 2712 RNA-related InterPro IDs was filtered to keep only 1029 “Family” IDs. The same procedure as used for the selection of the RBDs was applied to select InterPro “Family” IDs enriched in RBPs (Figure 12). This resulted in a final list of 672 RNA-related “Family” IDs, or Rfam IDs (Table 7). The proteins were then analyzed for the presence of these Rfam IDs.

The results of these analyses were used to produce Figures 13 to 16.

5.6. Retrieval of the MobiDB-lite data for disordered regions

Various algorithms to predict the presence of disorder in proteins exist today, with 45 of them being reviewed in 2019 (173). Several databases, such as Disprot (174), IDEAL (175), MobiDB (143) and D²P² (176), group the results of several of these algorithms, as well as some experimental data. I chose the MobiDB-lite prediction as it is more conservative, through the combination of ten different prediction tools (142, 147). Furthermore, it also includes experimental data from PDB in the prediction process, and its results are displayed as a consensus on widely used platforms like InterPro (145) and Uniprot (144). This algorithm provides the coordinates of each IDR in each protein, as well as the percentage of the protein length covered by IDRs, or disordered fraction of the protein. To retrieve the information for all the species of the RBP2GO database, the results of the MobiDB-lite algorithm were downloaded from the MobiDB database (<https://mobidb.bio.unipd.it>, Version 5.0.0) (142, 143) using the “download.file” function in R and the taxon ID of each species. Data could be retrieved for each species, but only 308 and 8 proteins were found to be annotated in Ec and ST respectively. These species were thus not displayed on the figures with IDR data.

The results of these analyses were used to produce Figures 17 to 19, and the data for each protein is present in the Table 8, along with the data generated in the previous sections.

5.7. GO enrichment analysis for non-RBPs containing an RBD

For each species, the GO enrichment analysis was performed in the non-RBPs containing at least one of the selected RBDs against the proteome with the PANTHER classification system (<http://pantherdb.org>, PANTHER 17.0 release from March 2021) accessed via the “rbioapi”

package in R (177) with a cut-off of 0.05 applied to the FDR (false discovery rate) values. The results for the different species were then merged in a single table. The GO terms were filtered to keep the terms with a p-value lower than 0.05. For each GO term, the number of species in which it was significantly enriched and its mean fold-enrichment in these species were calculated. Only the GO terms enriched in more than four species out of 13 and with a mean fold-enrichment over five were selected (Tables 9 and 10). The heatmaps of Figures 22B and 22C were created with the “geom_tile” function from the ggplot2 package (178), and the GO terms were classified by decreasing mean fold-enrichment.

5.8. Identification of new RBDs using the human proteins

Only the human proteins were included in this part of the analysis. To find new RBD candidates, I decided to take as my dataset of interest the RBPs with no RBD and with the top 20% highest RBP2GO scores, which represents 935 “high score” proteins. Three reference datasets were then constituted to detect the enriched InterPro IDs: 1 - the whole human proteome (20751 proteins), 2 - the non-RBPs without RBD (14411 proteins) and 3 - the RBPs without RBD and the 20% bottom lowest RBP2GO score (989 “low score” proteins). The InterPro IDs present in the “high score” dataset were extracted and the IDs that were present in our initial RBDs list were filtered out. Next, only the InterPro IDs of the type “Domain” or “Repeat” were taken into account, which resulted in a list of 637 IDs. For each InterPro ID (domain), I calculated the proportion of proteins containing the domain in the dataset of interest as well as in the three reference datasets. Next, for each domain, I computed its fold-enrichment in the dataset of interest as compared to each of the reference dataset. A p-value was also generated as a result of a Fisher’s exact test. The p-values were further adjusted for multiple testing using the FDR method (1). To select the most promising domains, I applied the following criteria: i) the domains that were significantly enriched against at least one reference dataset (more than a 2-fold enrichment and adjusted p-value less than 0.05) were selected, ii) the domain had to be present in at least 0.5% of the proteins of the “high score” dataset (in at least five proteins) and iii) each selected domain had an RBP vs non-RBPs hit ratio over one (same criteria as the one used in the first selection step of the RBDs). Altogether, 15 domains satisfied these three criteria (Table 11).

The results of these analyses were used to produce the Figures 21 and 22.

5.9. Validation of the newly discovered RBDs using published lists of RNA-binding peptides

To validate these newly discovered RBDs, I decided to take advantage of already published experimental data. Several studies compiled in RBP2GO used techniques allowing to identify the RNA-binding region of the detected RBP candidates (75–77, 83, 84, 123, 180, 181). The lists of the RNA-binding peptides identified in these studies were downloaded, keeping only the lists containing more than 1000 different peptides. They were then combined into a new dataset containing the peptide information with the Uniprot ID of the proteins they belong to and the coordinates of the peptides in each protein. Since four studies with more than 1000 peptides were found in human (75, 76, 83, 123), and only one each in mouse (180) and in *At* (181), I only considered the data for the human proteins. For each of the 624 proteins containing only the new RBDs (without containing any of the previously selected or non-selected RBDs), the proportion of peptides overlapping at least at 50% with the new RBD relative to the total number of peptides found in the protein was calculated. This number was normalized to the fraction of the protein covered by the domain. This number was then summed up for all the domains corresponding to one InterPro ID over all proteins containing at least one peptide, and normalized by the total number of occurrences of this domain in the proteins containing at least one peptide (Figure 23A). The same procedure was repeated for the selected RBDs (from Figure 5A) and for the non-selected domains (from Figure 5A) using groups of proteins containing only selected RBDs and only non-selected domains, respectively (Figure 23A). To stay in line with the selection criteria of the new RBDs, only the selected and non-selected RBDs that were present in at least five RBPs were taken into account. When taking into account the domains present in proteins with only one type of RBD, this resulted in two groups of 17 non-selected domains and 104 selected domains considered in this analysis.

The results of this validation can be found in Tables 1 and 7, and were used to produce Figure 23B.

5.10. Computation of the new RBP2GO composite score

To reflect the importance of the presence of RBDs or the relation to an Rfam ID for an RBP candidate, a new score was created, based on the previous RBP2GO score (108). The new “RBP2GO composite score” integrated the information already provided by the RBP2GO score and the knowledge on RBDs and Rfam IDs. It comprised three components. Component

1 represented one half of the score and was constituted by the listing count of the protein (in how many datasets is it detected as RBP candidate) divided by the number of datasets available for the species of the given protein. Component 1 was normalized and amounted to a maximum of 50. Component 2 represented one quarter of the score and was the mean of component 1 for the 10 first String interactors of the given protein (182). Component 2 was then normalized and amounted to a maximum of 25. The last quarter was constituted by component 3, a score pertaining to the presence of RBDs or Rfam IDs for the protein of interest.

For the component 3, a quality factor was attributed to each RBD/Rfam ID. To this aim, the ratio of hits in RBPs versus hits in non-RBPs in Hs was taken into consideration. If the ID was not found in human (Hs), the ratio in Mm was considered; if it was not found in mouse (Mm), the one in yeast was considered, and so on, following the order of the species by decreasing number of datasets (Figure 6). For each ID, this ratio was used to attribute a quality factor: if the ratio was less than 2, this factor will be 1, 2 if the ratio was above or equal to 2, 3 if the ratio was above or equal to 4, 4 if the ratio was above or equal to 8 and 5 for a ratio above or equal to 16. For infinite ratios (InterPro ID only present in RBPs), the number of RBPs with this ID was taken into account: for a number of RBPs less than or equal to 2, the quality factor was 2, 3 for a number of RBPs between 2 and 4, 4 for a number of RBPs between 4 and 8 and 5 for more than 8 RBPs. The attribution rules are summarized in the Table 3. Finally, the quality factors for all RBDs and RNA-related family IDs present for one protein were summed up, and this number was limited to 25. All the quality factors for the selected RBDs and Rfam IDs can be found in Table 12.

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

The Pearson correlation coefficients between the three components of the new composite score and the RBP2GO score were computed using the “cor” function of the “stats” package in R (183). The results are displayed in the Figure 24C.

The new RBP2GO composite score was analyzed in the four groups of proteins: non-RBPs without RBD nor Rfam ID, non-RBPs with RBD or Rfam ID, RBPs without RBD nor Rfam ID and RBP with RBD or Rfam ID to produce Figure 24B.

5.11. Selection of a list of high-confidence RBPs in Hs

Since more than 29% of all human proteins have been detected at least once in a proteome-wide screen compiled in the RBP2GO database (6100 proteins), it seemed necessary to take advantage of the knowledge gained in this study, in addition to the experimental data already compiled on RBP2GO, to establish a list of high-confidence human RBPs. Therefore, I took advantage of the new RBP2GO composite score, as it combines all of the knowledge acquired here and in the RBP2GO database (108). This list is comprised of all human proteins, whether they were detected as RBPs in proteome-wide screens or not, that have a composite score above or equal to 10, and at least two out of the three components of the score larger than zero. This resulted in a list of 2019 proteins, among which 1979 were already detected as RBPs in the screens of the RBP2GO database (available on the RBP2GO database: <https://rbp2go.dkfz.de>).

5.12. Update of the RBP2GO database

The lists of selected RBDs, completed with the 15 newly identified RBDs, and the list of selected Rfam IDs, are available on the RBP2GO database, in the “Download” section, along with the information on RBPs and non-RBPs for all 13 species. The RBD and Rfam status (does this protein contain an RBD or is annotated with an Rfam ID or not?) of each protein were added to these tables, along with the number and content fraction in RBDs, lists of present RBDs and Uniprot IDs, the presence of IDRs and disordered fraction for each protein. The RBD status, number of RBDs, RBD content fraction and list of present RBDs have been updated to take into account the 15 newly discovered RBDs.

All of this new information was also integrated into the RBP2GO database in a user friendly and intuitive manner. When searching for a protein in the “Protein Search” of each species, the results are now classified according to their RBP2GO composite score. This score has also been added to the “Protein Information” tab. A new tab has been integrated in the results of the “Protein Search” called “Domain Information” and displays all information on RBDs, Rfam IDs and IDRs for this selected protein. It also contains two tables that detail the present InterPro annotations for this protein, and if they are classified as RNA-binding or not in our study. It also allows the user to quickly have access to exterior information with multiple links towards the InterPro and the MobiDB databases (143, 145).

New options are also available in the “Advanced Search” tab. The users can now restrict their search to all protein containing a certain list of InterPro IDs, containing an RBD and/or containing an Rfam ID. The results can also be restricted to proteins with a certain minimum and/or maximum RBP2GO composite score.

Finally, a new section called “SEARCH DOMAINS” is present in the side bar of the database. This section presents three new tabs. The “RBD information” and “Rfam ID information” tabs respectively give information on the selection process of RBD and Rfam IDs, as well as a view of their distribution in the proteins of the database. The “Domain search” tab grants the users the possibility to search for a list of InterPro IDs and gives back a table with the InterPro ID, its type, its name, and its RNA-binding status (e.g., is it present in our list of RBDs and Rfam IDs or not).

6. Acknowledgements

First, I would like to thank Prof. Dr. Sven Diederichs for giving me the opportunity to work in his lab, first as a trainee and then to continue as a PhD student. I am grateful for his support throughout this project, and I would particularly like to express my gratitude for helping me through the difficult moments of the last years and allow me to explore different parts of the work as a researcher in biology. I also would like to thank PD Dr. Maiwen Caudron-Herger, who greatly contributed to my supervision and crafted most of the projects I worked on, as well as provided me with a lot of input and feedback over the years. Next, I would like to thank Prof. Dr. Karsten Rippe and Prof. Dr. Elmar Schiebel for their feedback and support as members of my thesis committee. I would also like to thank Prof. Dr. Oliver Gruß for his input as a member of my TAC committee.

Additionally, I would like to thank all of the B150 members that I had the chance to work with over the years, for their input, feedback and great support. A particular thought goes to Jeanette Seiler, who not only was always present to answer my questions related to science, but was also a great friend and mentor. I also would like to thank Varshni Rajagopal and Dr. Simona Cantarella, with whom I performed numerous cell cycle synchronizations, and who were always there to support and help me. I also would like to thank my collaborators from the Light Microscopy Facility at the DKFZ, Manuela Brom and Dr. Felix Bestvater.

I would also like to thank the friends I made here at the DKFZ, Sheldon, Francesco, Dwain, Angelika, Marta, Katharina, Ralf, with whom I have amazing memories of time spent together.

Finally, I express my gratitude to my family for their amazing support and their love. I am grateful to my parents and my sister, who saw me move away from them but still supported me in my career choices. I especially thank them for these hours on the phone spent to talk about my life in the lab. I am also very grateful to my parents in law, for accepting me and supporting me as their own daughter. And I owe a huge thank you to my husband. He joined forces with me, especially through the toughest moments, and I could not have accomplished half of what I did without him. His unconditional love has been and will always be my rock. Lastly, I would like to thank my son, Eden. I love you more than I could imagine, you are my little paradise.

7. Appendix

Table 9: Results of the GO term enrichment analysis for the biological process terms, limited to terms enriched in at least five species

Species	GO term label	Fold enrichment	Enriched in	Average enrichment
Hs	protein modification by small protein removal	7.07	5	5.02
Mm	protein modification by small protein removal	5.48	5	5.02
Sc	protein modification by small protein removal		5	5.02
Dm	protein modification by small protein removal	4.02	5	5.02
At	protein modification by small protein removal	2.61	5	5.02
Ce	protein modification by small protein removal	6.42	5	5.02
Dr	protein modification by small protein removal	3.22	5	5.02
Ec	protein modification by small protein removal		5	5.02
Pf	protein modification by small protein removal	8.50	5	5.02
Tb	protein modification by small protein removal	2.85	5	5.02
Hs	rRNA processing		5	5.05
Mm	rRNA processing	3.39	5	5.05
Sc	rRNA processing		5	5.05
Dm	rRNA processing	2.78	5	5.05
At	rRNA processing	3.15	5	5.05
Ce	rRNA processing	7.01	5	5.05
Dr	rRNA processing	7.60	5	5.05
Ec	rRNA processing		5	5.05
Pf	rRNA processing		5	5.05
Tb	rRNA processing	6.39	5	5.05
Hs	regulation of DNA metabolic process	3.70	7	5.13
Mm	regulation of DNA metabolic process	3.25	7	5.13
Sc	regulation of DNA metabolic process	4.46	7	5.13
Dm	regulation of DNA metabolic process	6.84	7	5.13
At	regulation of DNA metabolic process	3.33	7	5.13
Ce	regulation of DNA metabolic process	6.82	7	5.13
Dr	regulation of DNA metabolic process	4.72	7	5.13
Ec	regulation of DNA metabolic process		7	5.13
Pf	regulation of DNA metabolic process	9.92	7	5.13
Tb	regulation of DNA metabolic process	3.14	7	5.13
Hs	RNA metabolic process		7	5.23
Mm	RNA metabolic process	4.08	7	5.23
Sc	RNA metabolic process	2.03	7	5.23
Dm	RNA metabolic process	4.85	7	5.23
At	RNA metabolic process	4.87	7	5.23
Ce	RNA metabolic process	9.10	7	5.23
Dr	RNA metabolic process	7.55	7	5.23
Ec	RNA metabolic process	2.55	7	5.23
Pf	RNA metabolic process		7	5.23
Tb	RNA metabolic process	6.83	7	5.23
Hs	chromatin organization	6.33	7	5.31
Mm	chromatin organization	4.55	7	5.31
Sc	chromatin organization	4.08	7	5.31
Dm	chromatin organization	4.92	7	5.31

At	chromatin organization	4.35	7	5.31
Ce	chromatin organization	7.95	7	5.31
Dr	chromatin organization	5.66	7	5.31
Ec	chromatin organization		7	5.31
Pf	chromatin organization	7.44	7	5.31
Tb	chromatin organization	2.51	7	5.31
Hs	protein acetylation	8.99	5	5.40
Mm	protein acetylation	4.67	5	5.40
Sc	protein acetylation	3.59	5	5.40
Dm	protein acetylation	3.70	5	5.40
At	protein acetylation	2.55	5	5.40
Ce	protein acetylation	10.23	5	5.40
Dr	protein acetylation	4.05	5	5.40
Ec	protein acetylation		5	5.40
Pf	protein acetylation		5	5.40
Tb	protein acetylation		5	5.40
Hs	RNA processing		7	5.42
Mm	RNA processing	4.42	7	5.42
Sc	RNA processing	1.89	7	5.42
Dm	RNA processing	4.62	7	5.42
At	RNA processing	4.29	7	5.42
Ce	RNA processing	9.41	7	5.42
Dr	RNA processing	8.38	7	5.42
Ec	RNA processing	3.71	7	5.42
Pf	RNA processing		7	5.42
Tb	RNA processing	6.62	7	5.42
Hs	protein-DNA complex subunit organization		5	5.42
Mm	protein-DNA complex subunit organization	3.31	5	5.42
Sc	protein-DNA complex subunit organization	2.67	5	5.42
Dm	protein-DNA complex subunit organization	6.26	5	5.42
At	protein-DNA complex subunit organization	4.83	5	5.42
Ce	protein-DNA complex subunit organization	7.22	5	5.42
Dr	protein-DNA complex subunit organization	5.12	5	5.42
Ec	protein-DNA complex subunit organization		5	5.42
Pf	protein-DNA complex subunit organization	8.50	5	5.42
Tb	protein-DNA complex subunit organization		5	5.42
Hs	rRNA metabolic process		6	5.49
Mm	rRNA metabolic process	4.59	6	5.49
Sc	rRNA metabolic process		6	5.49
Dm	rRNA metabolic process	2.98	6	5.49
At	rRNA metabolic process	3.40	6	5.49
Ce	rRNA metabolic process	7.39	6	5.49
Dr	rRNA metabolic process	7.96	6	5.49
Ec	rRNA metabolic process		6	5.49
Pf	rRNA metabolic process		6	5.49
Tb	rRNA metabolic process	6.63	6	5.49
Hs	regulation of mRNA metabolic process		6	5.54
Mm	regulation of mRNA metabolic process	2.80	6	5.54
Sc	regulation of mRNA metabolic process		6	5.54
Dm	regulation of mRNA metabolic process	7.75	6	5.54
At	regulation of mRNA metabolic process	3.72	6	5.54
Ce	regulation of mRNA metabolic process	6.04	6	5.54
Dr	regulation of mRNA metabolic process	8.56	6	5.54

Ec	regulation of mRNA metabolic process		6	5.54
Pf	regulation of mRNA metabolic process		6	5.54
Tb	regulation of mRNA metabolic process	4.35	6	5.54
Hs	positive regulation of mRNA metabolic process		5	5.55
Mm	positive regulation of mRNA metabolic process	3.79	5	5.55
Sc	positive regulation of mRNA metabolic process		5	5.55
Dm	positive regulation of mRNA metabolic process	6.17	5	5.55
At	positive regulation of mRNA metabolic process	2.97	5	5.55
Ce	positive regulation of mRNA metabolic process	5.28	5	5.55
Dr	positive regulation of mRNA metabolic process	8.83	5	5.55
Ec	positive regulation of mRNA metabolic process		5	5.55
Pf	positive regulation of mRNA metabolic process		5	5.55
Tb	positive regulation of mRNA metabolic process	6.28	5	5.55
Hs	ncRNA processing		6	5.57
Mm	ncRNA processing	4.96	6	5.57
Sc	ncRNA processing		6	5.57
Dm	ncRNA processing	3.81	6	5.57
At	ncRNA processing	3.89	6	5.57
Ce	ncRNA processing	9.36	6	5.57
Dr	ncRNA processing	7.54	6	5.57
Ec	ncRNA processing	3.23	6	5.57
Pf	ncRNA processing		6	5.57
Tb	ncRNA processing	6.17	6	5.57
Hs	RNA splicing, via transesterification reactions		6	5.57
Mm	RNA splicing, via transesterification reactions	3.28	6	5.57
Sc	RNA splicing, via transesterification reactions	2.94	6	5.57
Dm	RNA splicing, via transesterification reactions	4.22	6	5.57
At	RNA splicing, via transesterification reactions	3.72	6	5.57
Ce	RNA splicing, via transesterification reactions	7.55	6	5.57
Dr	RNA splicing, via transesterification reactions	8.81	6	5.57
Ec	RNA splicing, via transesterification reactions		6	5.57
Pf	RNA splicing, via transesterification reactions		6	5.57
Tb	RNA splicing, via transesterification reactions	8.45	6	5.57
Hs	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile		6	5.59
Mm	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	3.28	6	5.59
Sc	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	3.09	6	5.59
Dm	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	4.26	6	5.59
At	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	3.72	6	5.59
Ce	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	7.55	6	5.59

Dr	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	8.81	6	5.59
Ec	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile		6	5.59
Pf	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile		6	5.59
Tb	RNA splicing, via transesterification reactions with bulged adenosine as nucleophile	8.45	6	5.59
Hs	mRNA splicing, via spliceosome		6	5.62
Mm	mRNA splicing, via spliceosome	3.28	6	5.62
Sc	mRNA splicing, via spliceosome	3.12	6	5.62
Dm	mRNA splicing, via spliceosome	4.26	6	5.62
At	mRNA splicing, via spliceosome	3.87	6	5.62
Ce	mRNA splicing, via spliceosome	7.55	6	5.62
Dr	mRNA splicing, via spliceosome	8.81	6	5.62
Ec	mRNA splicing, via spliceosome		6	5.62
Pf	mRNA splicing, via spliceosome		6	5.62
Tb	mRNA splicing, via spliceosome	8.45	6	5.62
Hs	DNA-templated transcription initiation		6	5.75
Mm	DNA-templated transcription initiation	3.31	6	5.75
Sc	DNA-templated transcription initiation	5.02	6	5.75
Dm	DNA-templated transcription initiation	4.45	6	5.75
At	DNA-templated transcription initiation	4.46	6	5.75
Ce	DNA-templated transcription initiation	9.74	6	5.75
Dr	DNA-templated transcription initiation	4.67	6	5.75
Ec	DNA-templated transcription initiation		6	5.75
Pf	DNA-templated transcription initiation	5.95	6	5.75
Tb	DNA-templated transcription initiation	8.37	6	5.75
Hs	RNA splicing		7	5.77
Mm	RNA splicing	3.76	7	5.77
Sc	RNA splicing	3.81	7	5.77
Dm	RNA splicing	4.66	7	5.77
At	RNA splicing	4.23	7	5.77
Ce	RNA splicing	7.82	7	5.77
Dr	RNA splicing	8.69	7	5.77
Ec	RNA splicing		7	5.77
Pf	RNA splicing		7	5.77
Tb	RNA splicing	7.42	7	5.77
Hs	ncRNA metabolic process		6	5.78
Mm	ncRNA metabolic process	5.86	6	5.78
Sc	ncRNA metabolic process	1.83	6	5.78
Dm	ncRNA metabolic process	4.78	6	5.78
At	ncRNA metabolic process	4.32	6	5.78
Ce	ncRNA metabolic process	8.88	6	5.78
Dr	ncRNA metabolic process	8.08	6	5.78
Ec	ncRNA metabolic process		6	5.78
Pf	ncRNA metabolic process		6	5.78
Tb	ncRNA metabolic process	6.70	6	5.78
Hs	protein deubiquitination	7.92	6	5.92

Mm	protein deubiquitination	6.46	6	5.92
Sc	protein deubiquitination		6	5.92
Dm	protein deubiquitination	5.39	6	5.92
At	protein deubiquitination	3.28	6	5.92
Ce	protein deubiquitination	7.79	6	5.92
Dr	protein deubiquitination	3.78	6	5.92
Ec	protein deubiquitination		6	5.92
Pf	protein deubiquitination	9.40	6	5.92
Tb	protein deubiquitination	3.31	6	5.92
Hs	mRNA metabolic process		6	5.98
Mm	mRNA metabolic process	3.74	6	5.98
Sc	mRNA metabolic process	2.04	6	5.98
Dm	mRNA metabolic process	5.38	6	5.98
At	mRNA metabolic process	4.35	6	5.98
Ce	mRNA metabolic process	9.42	6	5.98
Dr	mRNA metabolic process	9.08	6	5.98
Ec	mRNA metabolic process		6	5.98
Pf	mRNA metabolic process		6	5.98
Tb	mRNA metabolic process	7.85	6	5.98
Hs	tRNA processing		6	6.03
Mm	tRNA processing	7.45	6	6.03
Sc	tRNA processing	2.43	6	6.03
Dm	tRNA processing	4.19	6	6.03
At	tRNA processing	4.54	6	6.03
Ce	tRNA processing	12.51	6	6.03
Dr	tRNA processing	6.76	6	6.03
Ec	tRNA processing	5.00	6	6.03
Pf	tRNA processing		6	6.03
Tb	tRNA processing	5.35	6	6.03
Hs	mRNA processing		7	6.08
Mm	mRNA processing	4.03	7	6.08
Sc	mRNA processing	2.83	7	6.08
Dm	mRNA processing	5.01	7	6.08
At	mRNA processing	4.50	7	6.08
Ce	mRNA processing	9.38	7	6.08
Dr	mRNA processing	9.02	7	6.08
Ec	mRNA processing		7	6.08
Pf	mRNA processing		7	6.08
Tb	mRNA processing	7.78	7	6.08
Hs	mRNA catabolic process		5	6.10
Mm	mRNA catabolic process	3.74	5	6.10
Sc	mRNA catabolic process		5	6.10
Dm	mRNA catabolic process	6.03	5	6.10
At	mRNA catabolic process	2.37	5	6.10
Ce	mRNA catabolic process	8.07	5	6.10
Dr	mRNA catabolic process	9.08	5	6.10
Ec	mRNA catabolic process		5	6.10
Pf	mRNA catabolic process		5	6.10
Tb	mRNA catabolic process	7.29	5	6.10
Hs	peptidyl-lysine acetylation	10.34	6	6.28
Mm	peptidyl-lysine acetylation	5.31	6	6.28
Sc	peptidyl-lysine acetylation	4.28	6	6.28
Dm	peptidyl-lysine acetylation	4.02	6	6.28

At	peptidyl-lysine acetylation	2.93	6	6.28
Ce	peptidyl-lysine acetylation	12.37	6	6.28
Dr	peptidyl-lysine acetylation	4.69	6	6.28
Ec	peptidyl-lysine acetylation		6	6.28
Pf	peptidyl-lysine acetylation		6	6.28
Tb	peptidyl-lysine acetylation		6	6.28
Hs	internal protein amino acid acetylation	10.41	6	6.30
Mm	internal protein amino acid acetylation	5.43	6	6.30
Sc	internal protein amino acid acetylation	4.28	6	6.30
Dm	internal protein amino acid acetylation	4.02	6	6.30
At	internal protein amino acid acetylation	2.93	6	6.30
Ce	internal protein amino acid acetylation	12.37	6	6.30
Dr	internal protein amino acid acetylation	4.69	6	6.30
Ec	internal protein amino acid acetylation		6	6.30
Pf	internal protein amino acid acetylation		6	6.30
Tb	internal protein amino acid acetylation		6	6.30
Hs	mitochondrial translation		6	6.31
Mm	mitochondrial translation	3.65	6	6.31
Sc	mitochondrial translation	3.48	6	6.31
Dm	mitochondrial translation		6	6.31
At	mitochondrial translation	5.27	6	6.31
Ce	mitochondrial translation	5.45	6	6.31
Dr	mitochondrial translation	9.75	6	6.31
Ec	mitochondrial translation		6	6.31
Pf	mitochondrial translation		6	6.31
Tb	mitochondrial translation	10.28	6	6.31
Hs	internal peptidyl-lysine acetylation	10.57	6	6.33
Mm	internal peptidyl-lysine acetylation	5.43	6	6.33
Sc	internal peptidyl-lysine acetylation	4.28	6	6.33
Dm	internal peptidyl-lysine acetylation	4.02	6	6.33
At	internal peptidyl-lysine acetylation	2.93	6	6.33
Ce	internal peptidyl-lysine acetylation	12.37	6	6.33
Dr	internal peptidyl-lysine acetylation	4.69	6	6.33
Ec	internal peptidyl-lysine acetylation		6	6.33
Pf	internal peptidyl-lysine acetylation		6	6.33
Tb	internal peptidyl-lysine acetylation		6	6.33
Hs	tRNA metabolic process		7	6.34
Mm	tRNA metabolic process	7.89	7	6.34
Sc	tRNA metabolic process	2.67	7	6.34
Dm	tRNA metabolic process	5.75	7	6.34
At	tRNA metabolic process	5.38	7	6.34
Ce	tRNA metabolic process	9.92	7	6.34
Dr	tRNA metabolic process	8.34	7	6.34
Ec	tRNA metabolic process	3.65	7	6.34
Pf	tRNA metabolic process		7	6.34
Tb	tRNA metabolic process	7.07	7	6.34
Hs	nuclear-transcribed mRNA catabolic process		5	6.36
Mm	nuclear-transcribed mRNA catabolic process	4.10	5	6.36
Sc	nuclear-transcribed mRNA catabolic process		5	6.36
Dm	nuclear-transcribed mRNA catabolic process	4.82	5	6.36
At	nuclear-transcribed mRNA catabolic process	2.43	5	6.36
Ce	nuclear-transcribed mRNA catabolic process	8.89	5	6.36
Dr	nuclear-transcribed mRNA catabolic process	10.03	5	6.36

Ec	nuclear-transcribed mRNA catabolic process		5	6.36
Pf	nuclear-transcribed mRNA catabolic process		5	6.36
Tb	nuclear-transcribed mRNA catabolic process	7.91	5	6.36
Hs	nucleosome organization	4.02	5	6.44
Mm	nucleosome organization	4.68	5	6.44
Sc	nucleosome organization	3.71	5	6.44
Dm	nucleosome organization	8.71	5	6.44
At	nucleosome organization	6.06	5	6.44
Ce	nucleosome organization	9.88	5	6.44
Dr	nucleosome organization	7.10	5	6.44
Ec	nucleosome organization		5	6.44
Pf	nucleosome organization	11.16	5	6.44
Tb	nucleosome organization	2.69	5	6.44
Hs	tRNA modification		5	6.52
Mm	tRNA modification	7.18	5	6.52
Sc	tRNA modification		5	6.52
Dm	tRNA modification		5	6.52
At	tRNA modification	2.86	5	6.52
Ce	tRNA modification	11.04	5	6.52
Dr	tRNA modification	6.27	5	6.52
Ec	tRNA modification	6.19	5	6.52
Pf	tRNA modification		5	6.52
Tb	tRNA modification	5.55	5	6.52
Hs	histone acetylation	11.08	6	6.60
Mm	histone acetylation	5.84	6	6.60
Sc	histone acetylation	4.36	6	6.60
Dm	histone acetylation	4.12	6	6.60
At	histone acetylation	2.93	6	6.60
Ce	histone acetylation	12.97	6	6.60
Dr	histone acetylation	4.86	6	6.60
Ec	histone acetylation		6	6.60
Pf	histone acetylation		6	6.60
Tb	histone acetylation		6	6.60
Hs	RNA catabolic process		6	6.68
Mm	RNA catabolic process	4.49	6	6.68
Sc	RNA catabolic process		6	6.68
Dm	RNA catabolic process	5.89	6	6.68
At	RNA catabolic process	2.93	6	6.68
Ce	RNA catabolic process	9.28	6	6.68
Dr	RNA catabolic process	9.20	6	6.68
Ec	RNA catabolic process		6	6.68
Pf	RNA catabolic process		6	6.68
Tb	RNA catabolic process	8.27	6	6.68
Hs	regulation of RNA splicing		5	6.69
Mm	regulation of RNA splicing	3.08	5	6.69
Sc	regulation of RNA splicing		5	6.69
Dm	regulation of RNA splicing	8.45	5	6.69
At	regulation of RNA splicing	5.58	5	6.69
Ce	regulation of RNA splicing	7.31	5	6.69
Dr	regulation of RNA splicing	9.05	5	6.69
Ec	regulation of RNA splicing		5	6.69
Pf	regulation of RNA splicing		5	6.69
Tb	regulation of RNA splicing		5	6.69

Hs	RNA modification		5	6.79
Mm	RNA modification	5.78	5	6.79
Sc	RNA modification		5	6.79
Dm	RNA modification	2.65	5	6.79
At	RNA modification	7.59	5	6.79
Ce	RNA modification	12.19	5	6.79
Dr	RNA modification	7.65	5	6.79
Ec	RNA modification	4.71	5	6.79
Pf	RNA modification		5	6.79
Tb	RNA modification	6.93	5	6.79
Hs	mitochondrial gene expression		6	6.83
Mm	mitochondrial gene expression	4.26	6	6.83
Sc	mitochondrial gene expression	4.04	6	6.83
Dm	mitochondrial gene expression		6	6.83
At	mitochondrial gene expression	7.79	6	6.83
Ce	mitochondrial gene expression	8.77	6	6.83
Dr	mitochondrial gene expression	9.54	6	6.83
Ec	mitochondrial gene expression		6	6.83
Pf	mitochondrial gene expression		6	6.83
Tb	mitochondrial gene expression	6.61	6	6.83
Hs	nucleic acid phosphodiester bond hydrolysis		6	6.83
Mm	nucleic acid phosphodiester bond hydrolysis	7.85	6	6.83
Sc	nucleic acid phosphodiester bond hydrolysis		6	6.83
Dm	nucleic acid phosphodiester bond hydrolysis	3.72	6	6.83
At	nucleic acid phosphodiester bond hydrolysis	4.71	6	6.83
Ce	nucleic acid phosphodiester bond hydrolysis	8.73	6	6.83
Dr	nucleic acid phosphodiester bond hydrolysis	7.32	6	6.83
Ec	nucleic acid phosphodiester bond hydrolysis	8.72	6	6.83
Pf	nucleic acid phosphodiester bond hydrolysis		6	6.83
Tb	nucleic acid phosphodiester bond hydrolysis	6.79	6	6.83
Hs	nucleosome assembly		5	7.20
Mm	nucleosome assembly	4.26	5	7.20
Sc	nucleosome assembly		5	7.20
Dm	nucleosome assembly	7.48	5	7.20
At	nucleosome assembly	5.83	5	7.20
Ce	nucleosome assembly	11.16	5	7.20
Dr	nucleosome assembly	7.29	5	7.20
Ec	nucleosome assembly		5	7.20
Pf	nucleosome assembly		5	7.20
Tb	nucleosome assembly		5	7.20
Hs	DNA conformation change	5.55	5	7.28
Mm	DNA conformation change	5.14	5	7.28
Sc	DNA conformation change	11.92	5	7.28
Dm	DNA conformation change		5	7.28
At	DNA conformation change	5.39	5	7.28
Ce	DNA conformation change		5	7.28
Dr	DNA conformation change	4.45	5	7.28
Ec	DNA conformation change	15.88	5	7.28
Pf	DNA conformation change		5	7.28
Tb	DNA conformation change	2.62	5	7.28
Hs	mRNA 3'-end processing		6	7.54
Mm	mRNA 3'-end processing	5.76	6	7.54
Sc	mRNA 3'-end processing		6	7.54

Dm	mRNA 3'-end processing	5.86	6	7.54
At	mRNA 3'-end processing	4.07	6	7.54
Ce	mRNA 3'-end processing	16.36	6	7.54
Dr	mRNA 3'-end processing	8.19	6	7.54
Ec	mRNA 3'-end processing		6	7.54
Pf	mRNA 3'-end processing		6	7.54
Tb	mRNA 3'-end processing	5.02	6	7.54
Hs	RNA phosphodiester bond hydrolysis		5	8.02
Mm	RNA phosphodiester bond hydrolysis	10.71	5	8.02
Sc	RNA phosphodiester bond hydrolysis		5	8.02
Dm	RNA phosphodiester bond hydrolysis	3.69	5	8.02
At	RNA phosphodiester bond hydrolysis	5.06	5	8.02
Ce	RNA phosphodiester bond hydrolysis	10.34	5	8.02
Dr	RNA phosphodiester bond hydrolysis	10.09	5	8.02
Ec	RNA phosphodiester bond hydrolysis		5	8.02
Pf	RNA phosphodiester bond hydrolysis		5	8.02
Tb	RNA phosphodiester bond hydrolysis	8.21	5	8.02
Hs	RNA 3'-end processing		6	8.03
Mm	RNA 3'-end processing	7.20	6	8.03
Sc	RNA 3'-end processing		6	8.03
Dm	RNA 3'-end processing	6.75	6	8.03
At	RNA 3'-end processing	5.27	6	8.03
Ce	RNA 3'-end processing	13.09	6	8.03
Dr	RNA 3'-end processing	9.20	6	8.03
Ec	RNA 3'-end processing		6	8.03
Pf	RNA 3'-end processing		6	8.03
Tb	RNA 3'-end processing	6.65	6	8.03
Hs	RNA methylation		5	8.03
Mm	RNA methylation	8.51	5	8.03
Sc	RNA methylation	3.39	5	8.03
Dm	RNA methylation		5	8.03
At	RNA methylation	4.74	5	8.03
Ce	RNA methylation	15.55	5	8.03
Dr	RNA methylation	8.16	5	8.03
Ec	RNA methylation		5	8.03
Pf	RNA methylation		5	8.03
Tb	RNA methylation	7.81	5	8.03
Hs	demethylation	10.82	5	8.10
Mm	demethylation	7.41	5	8.10
Sc	demethylation	9.27	5	8.10
Dm	demethylation		5	8.10
At	demethylation	4.35	5	8.10
Ce	demethylation	10.23	5	8.10
Dr	demethylation	6.25	5	8.10
Ec	demethylation		5	8.10
Pf	demethylation		5	8.10
Tb	demethylation	8.37	5	8.10
Hs	amino acid activation		6	8.12
Mm	amino acid activation	9.78	6	8.12
Sc	amino acid activation	3.48	6	8.12
Dm	amino acid activation	6.89	6	8.12
At	amino acid activation	7.91	6	8.12
Ce	amino acid activation	5.84	6	8.12

Dr	amino acid activation	11.88	6	8.12
Ec	amino acid activation		6	8.12
Pf	amino acid activation		6	8.12
Tb	amino acid activation	11.05	6	8.12
Hs	DNA geometric change	6.06	5	8.15
Mm	DNA geometric change	5.41	5	8.15
Sc	DNA geometric change	13.41	5	8.15
Dm	DNA geometric change	4.02	5	8.15
At	DNA geometric change	7.19	5	8.15
Ce	DNA geometric change		5	8.15
Dr	DNA geometric change	5.12	5	8.15
Ec	DNA geometric change	21.48	5	8.15
Pf	DNA geometric change		5	8.15
Tb	DNA geometric change	2.51	5	8.15
Hs	DNA duplex unwinding	6.45	5	8.25
Mm	DNA duplex unwinding	5.52	5	8.25
Sc	DNA duplex unwinding	13.41	5	8.25
Dm	DNA duplex unwinding	4.34	5	8.25
At	DNA duplex unwinding	7.19	5	8.25
Ce	DNA duplex unwinding		5	8.25
Dr	DNA duplex unwinding	5.12	5	8.25
Ec	DNA duplex unwinding	21.48	5	8.25
Pf	DNA duplex unwinding		5	8.25
Tb	DNA duplex unwinding	2.51	5	8.25
Hs	tRNA aminoacylation		6	8.28
Mm	tRNA aminoacylation	10.01	6	8.28
Sc	tRNA aminoacylation	3.48	6	8.28
Dm	tRNA aminoacylation	7.23	6	8.28
At	tRNA aminoacylation	7.91	6	8.28
Ce	tRNA aminoacylation	6.14	6	8.28
Dr	tRNA aminoacylation	12.12	6	8.28
Ec	tRNA aminoacylation		6	8.28
Pf	tRNA aminoacylation		6	8.28
Tb	tRNA aminoacylation	11.05	6	8.28
Hs	tRNA aminoacylation for protein translation		6	8.31
Mm	tRNA aminoacylation for protein translation	9.93	6	8.31
Sc	tRNA aminoacylation for protein translation	3.01	6	8.31
Dm	tRNA aminoacylation for protein translation	6.84	6	8.31
At	tRNA aminoacylation for protein translation	7.91	6	8.31
Ce	tRNA aminoacylation for protein translation	6.63	6	8.31
Dr	tRNA aminoacylation for protein translation	12.82	6	8.31
Ec	tRNA aminoacylation for protein translation		6	8.31
Pf	tRNA aminoacylation for protein translation		6	8.31
Tb	tRNA aminoacylation for protein translation	10.99	6	8.31
Hs	regulation of DNA recombination	5.33	5	8.39
Mm	regulation of DNA recombination	4.88	5	8.39
Sc	regulation of DNA recombination	5.06	5	8.39
Dm	regulation of DNA recombination	15.23	5	8.39
At	regulation of DNA recombination	7.91	5	8.39
Ce	regulation of DNA recombination	12.03	5	8.39
Dr	regulation of DNA recombination	8.29	5	8.39
Ec	regulation of DNA recombination		5	8.39
Pf	regulation of DNA recombination		5	8.39

Tb	regulation of DNA recombination		5	8.39
Hs	mitochondrial RNA metabolic process		5	9.28
Mm	mitochondrial RNA metabolic process	5.65	5	9.28
Sc	mitochondrial RNA metabolic process	6.55	5	9.28
Dm	mitochondrial RNA metabolic process		5	9.28
At	mitochondrial RNA metabolic process	8.36	5	9.28
Ce	mitochondrial RNA metabolic process	16.36	5	9.28
Dr	mitochondrial RNA metabolic process	9.50	5	9.28
Ec	mitochondrial RNA metabolic process		5	9.28
Pf	mitochondrial RNA metabolic process		5	9.28
Tb	mitochondrial RNA metabolic process		5	9.28
Hs	negative regulation of DNA metabolic process	4.68	5	9.29
Mm	negative regulation of DNA metabolic process	4.27	5	9.29
Sc	negative regulation of DNA metabolic process	3.57	5	9.29
Dm	negative regulation of DNA metabolic process	9.04	5	9.29
At	negative regulation of DNA metabolic process	6.89	5	9.29
Ce	negative regulation of DNA metabolic process	11.16	5	9.29
Dr	negative regulation of DNA metabolic process	7.97	5	9.29
Ec	negative regulation of DNA metabolic process		5	9.29
Pf	negative regulation of DNA metabolic process	29.76	5	9.29
Tb	negative regulation of DNA metabolic process	6.28	5	9.29
Hs	ncRNA catabolic process		5	9.77
Mm	ncRNA catabolic process	8.28	5	9.77
Sc	ncRNA catabolic process		5	9.77
Dm	ncRNA catabolic process	12.05	5	9.77
At	ncRNA catabolic process	4.61	5	9.77
Ce	ncRNA catabolic process		5	9.77
Dr	ncRNA catabolic process	12.50	5	9.77
Ec	ncRNA catabolic process		5	9.77
Pf	ncRNA catabolic process		5	9.77
Tb	ncRNA catabolic process	11.42	5	9.77
Hs	RNA polyadenylation		5	10.09
Mm	RNA polyadenylation	7.24	5	10.09
Sc	RNA polyadenylation		5	10.09
Dm	RNA polyadenylation	8.68	5	10.09
At	RNA polyadenylation	4.87	5	10.09
Ce	RNA polyadenylation	24.55	5	10.09
Dr	RNA polyadenylation	10.39	5	10.09
Ec	RNA polyadenylation		5	10.09
Pf	RNA polyadenylation		5	10.09
Tb	RNA polyadenylation	4.83	5	10.09
Hs	negative regulation of DNA recombination	6.60	5	11.17
Mm	negative regulation of DNA recombination	6.34	5	11.17
Sc	negative regulation of DNA recombination		5	11.17
Dm	negative regulation of DNA recombination	18.08	5	11.17
At	negative regulation of DNA recombination	9.16	5	11.17
Ce	negative regulation of DNA recombination	15.74	5	11.17
Dr	negative regulation of DNA recombination	11.13	5	11.17
Ec	negative regulation of DNA recombination		5	11.17
Pf	negative regulation of DNA recombination		5	11.17
Tb	negative regulation of DNA recombination		5	11.17
Hs	cell redox homeostasis	7.07	5	11.23
Mm	cell redox homeostasis		5	11.23

Sc	cell redox homeostasis		5	11.23
Dm	cell redox homeostasis	12.05	5	11.23
At	cell redox homeostasis	6.68	5	11.23
Ce	cell redox homeostasis	12.59	5	11.23
Dr	cell redox homeostasis	7.03	5	11.23
Ec	cell redox homeostasis		5	11.23
Pf	cell redox homeostasis	26.04	5	11.23
Tb	cell redox homeostasis	7.18	5	11.23
Hs	DNA-templated DNA replication maintenance of fidelity		6	14.89
Mm	DNA-templated DNA replication maintenance of fidelity	5.19	6	14.89
Sc	DNA-templated DNA replication maintenance of fidelity		6	14.89
Dm	DNA-templated DNA replication maintenance of fidelity		6	14.89
At	DNA-templated DNA replication maintenance of fidelity	7.38	6	14.89
Ce	DNA-templated DNA replication maintenance of fidelity	11.16	6	14.89
Dr	DNA-templated DNA replication maintenance of fidelity	5.14	6	14.89
Ec	DNA-templated DNA replication maintenance of fidelity	31.30	6	14.89
Pf	DNA-templated DNA replication maintenance of fidelity	35.71	6	14.89
Tb	DNA-templated DNA replication maintenance of fidelity	8.37	6	14.89
Hs	tRNA-type intron splice site recognition and cleavage		5	28.11
Mm	tRNA-type intron splice site recognition and cleavage	33.11	5	28.11
Sc	tRNA-type intron splice site recognition and cleavage	27.82	5	28.11
Dm	tRNA-type intron splice site recognition and cleavage	36.16	5	28.11
At	tRNA-type intron splice site recognition and cleavage	15.81	5	28.11
Ce	tRNA-type intron splice site recognition and cleavage	40.91	5	28.11
Dr	tRNA-type intron splice site recognition and cleavage	14.84	5	28.11
Ec	tRNA-type intron splice site recognition and cleavage		5	28.11
Pf	tRNA-type intron splice site recognition and cleavage		5	28.11
Tb	tRNA-type intron splice site recognition and cleavage		5	28.11

Table 10: Results of the GO term enrichment analysis for the molecular function terms, limited to the terms enriched in at least 5 species

Species	GO term label	Fold enrichment	Enriched in	Average enrichment
Hs	isomerase activity	5.96	8	5.09
Mm	isomerase activity	3.74	8	5.09
Sc	isomerase activity		8	5.09
Dm	isomerase activity	5.17	8	5.09
At	isomerase activity	2.84	8	5.09
Ce	isomerase activity	7.65	8	5.09
Dr	isomerase activity	4.75	8	5.09
Ec	isomerase activity		8	5.09
Pf	isomerase activity	5.95	8	5.09
Tb	isomerase activity	4.67	8	5.09
Hs	catalytic activity, acting on DNA	5.61	10	5.24
Mm	catalytic activity, acting on DNA	6.89	10	5.24
Sc	catalytic activity, acting on DNA	5.51	10	5.24
Dm	catalytic activity, acting on DNA	3.64	10	5.24
At	catalytic activity, acting on DNA	4.96	10	5.24
Ce	catalytic activity, acting on DNA	4.22	10	5.24
Dr	catalytic activity, acting on DNA	5.07	10	5.24
Ec	catalytic activity, acting on DNA	7.10	10	5.24
Pf	catalytic activity, acting on DNA	6.14	10	5.24
Tb	catalytic activity, acting on DNA	3.23	10	5.24
Hs	translation factor activity, RNA binding		5	5.50
Mm	translation factor activity, RNA binding	5.89	5	5.50
Sc	translation factor activity, RNA binding		5	5.50
Dm	translation factor activity, RNA binding	3.96	5	5.50
At	translation factor activity, RNA binding	1.69	5	5.50
Ce	translation factor activity, RNA binding	5.93	5	5.50
Dr	translation factor activity, RNA binding	8.46	5	5.50
Ec	translation factor activity, RNA binding		5	5.50
Pf	translation factor activity, RNA binding		5	5.50
Tb	translation factor activity, RNA binding	7.07	5	5.50
Hs	translation regulator activity		5	5.54
Mm	translation regulator activity	4.63	5	5.54
Sc	translation regulator activity		5	5.54
Dm	translation regulator activity	4.96	5	5.54
At	translation regulator activity	1.99	5	5.54
Ce	translation regulator activity	6.33	5	5.54
Dr	translation regulator activity	8.58	5	5.54
Ec	translation regulator activity		5	5.54
Pf	translation regulator activity		5	5.54
Tb	translation regulator activity	6.75	5	5.54
Hs	translation regulator activity, nucleic acid binding		5	5.69
Mm	translation regulator activity, nucleic acid binding	5.52	5	5.69
Sc	translation regulator activity, nucleic acid binding		5	5.69
Dm	translation regulator activity, nucleic acid binding	5.17	5	5.69

At	translation regulator activity, nucleic acid binding	1.66	5	5.69
Ce	translation regulator activity, nucleic acid binding	6.08	5	5.69
Dr	translation regulator activity, nucleic acid binding	8.65	5	5.69
Ec	translation regulator activity, nucleic acid binding		5	5.69
Pf	translation regulator activity, nucleic acid binding		5	5.69
Tb	translation regulator activity, nucleic acid binding	7.07	5	5.69
Hs	catalytic activity, acting on a nucleic acid	2.64	10	5.75
Mm	catalytic activity, acting on a nucleic acid	8.34	10	5.75
Sc	catalytic activity, acting on a nucleic acid	4.42	10	5.75
Dm	catalytic activity, acting on a nucleic acid	4.93	10	5.75
At	catalytic activity, acting on a nucleic acid	5.70	10	5.75
Ce	catalytic activity, acting on a nucleic acid	8.52	10	5.75
Dr	catalytic activity, acting on a nucleic acid	7.97	10	5.75
Ec	catalytic activity, acting on a nucleic acid	5.58	10	5.75
Pf	catalytic activity, acting on a nucleic acid	3.36	10	5.75
Tb	catalytic activity, acting on a nucleic acid	6.08	10	5.75
Hs	ubiquitin-like protein peptidase activity	9.06	7	5.86
Mm	ubiquitin-like protein peptidase activity	6.33	7	5.86
Sc	ubiquitin-like protein peptidase activity		7	5.86
Dm	ubiquitin-like protein peptidase activity	4.78	7	5.86
At	ubiquitin-like protein peptidase activity	3.13	7	5.86
Ce	ubiquitin-like protein peptidase activity	6.29	7	5.86
Dr	ubiquitin-like protein peptidase activity	3.64	7	5.86
Ec	ubiquitin-like protein peptidase activity		7	5.86
Pf	ubiquitin-like protein peptidase activity	9.92	7	5.86
Tb	ubiquitin-like protein peptidase activity	3.69	7	5.86
Hs	nucleotidyltransferase activity		7	6.14
Mm	nucleotidyltransferase activity	6.90	7	6.14
Sc	nucleotidyltransferase activity	2.21	7	6.14
Dm	nucleotidyltransferase activity	6.48	7	6.14
At	nucleotidyltransferase activity	4.26	7	6.14
Ce	nucleotidyltransferase activity	12.32	7	6.14
Dr	nucleotidyltransferase activity	5.32	7	6.14
Ec	nucleotidyltransferase activity		7	6.14
Pf	nucleotidyltransferase activity	6.13	7	6.14
Tb	nucleotidyltransferase activity	5.53	7	6.14
Hs	endonuclease activity		6	6.34
Mm	endonuclease activity	7.94	6	6.34
Sc	endonuclease activity	2.48	6	6.34
Dm	endonuclease activity	4.38	6	6.34
At	endonuclease activity	8.49	6	6.34
Ce	endonuclease activity	10.71	6	6.34
Dr	endonuclease activity	6.53	6	6.34
Ec	endonuclease activity	4.66	6	6.34
Pf	endonuclease activity		6	6.34
Tb	endonuclease activity	5.54	6	6.34
Hs	nuclease activity	3.25	8	6.50

Mm	nuclease activity	8.70	8	6.50
Sc	nuclease activity		8	6.50
Dm	nuclease activity	4.80	8	6.50
At	nuclease activity	7.08	8	6.50
Ce	nuclease activity	9.02	8	6.50
Dr	nuclease activity	7.33	8	6.50
Ec	nuclease activity	7.20	8	6.50
Pf	nuclease activity	5.47	8	6.50
Tb	nuclease activity	5.67	8	6.50
Hs	deubiquitinase activity	10.06	7	6.58
Mm	deubiquitinase activity	7.36	7	6.58
Sc	deubiquitinase activity		7	6.58
Dm	deubiquitinase activity	5.89	7	6.58
At	deubiquitinase activity	3.60	7	6.58
Ce	deubiquitinase activity	7.11	7	6.58
Dr	deubiquitinase activity	4.20	7	6.58
Ec	deubiquitinase activity		7	6.58
Pf	deubiquitinase activity	10.50	7	6.58
Tb	deubiquitinase activity	3.93	7	6.58
Hs	cis-trans isomerase activity	12.37	5	6.66
Mm	cis-trans isomerase activity	8.49	5	6.66
Sc	cis-trans isomerase activity		5	6.66
Dm	cis-trans isomerase activity	3.29	5	6.66
At	cis-trans isomerase activity	5.44	5	6.66
Ce	cis-trans isomerase activity	7.44	5	6.66
Dr	cis-trans isomerase activity	7.42	5	6.66
Ec	cis-trans isomerase activity		5	6.66
Pf	cis-trans isomerase activity	6.87	5	6.66
Tb	cis-trans isomerase activity	2.01	5	6.66
Hs	exonuclease activity	5.96	6	6.78
Mm	exonuclease activity	7.71	6	6.78
Sc	exonuclease activity		6	6.78
Dm	exonuclease activity	5.50	6	6.78
At	exonuclease activity	5.22	6	6.78
Ce	exonuclease activity	4.35	6	6.78
Dr	exonuclease activity	7.89	6	6.78
Ec	exonuclease activity	9.42	6	6.78
Pf	exonuclease activity	8.93	6	6.78
Tb	exonuclease activity	6.08	6	6.78
Hs	dioxygenase activity	8.51	6	6.82
Mm	dioxygenase activity	7.20	6	6.82
Sc	dioxygenase activity	5.96	6	6.82
Dm	dioxygenase activity		6	6.82
At	dioxygenase activity	9.04	6	6.82
Ce	dioxygenase activity	5.97	6	6.82
Dr	dioxygenase activity	3.64	6	6.82
Ec	dioxygenase activity		6	6.82
Pf	dioxygenase activity		6	6.82
Tb	dioxygenase activity	7.39	6	6.82
Hs	catalytic activity, acting on RNA		8	7.08
Mm	catalytic activity, acting on RNA	9.09	8	7.08
Sc	catalytic activity, acting on RNA	2.88	8	7.08
Dm	catalytic activity, acting on RNA	5.39	8	7.08

At	catalytic activity, acting on RNA	6.14	8	7.08
Ce	catalytic activity, acting on RNA	10.26	8	7.08
Dr	catalytic activity, acting on RNA	9.52	8	7.08
Ec	catalytic activity, acting on RNA	5.25	8	7.08
Pf	catalytic activity, acting on RNA		8	7.08
Tb	catalytic activity, acting on RNA	8.11	8	7.08
Hs	peptidyl-prolyl cis-trans isomerase activity	13.19	5	7.16
Mm	peptidyl-prolyl cis-trans isomerase activity	9.46	5	7.16
Sc	peptidyl-prolyl cis-trans isomerase activity		5	7.16
Dm	peptidyl-prolyl cis-trans isomerase activity	3.50	5	7.16
At	peptidyl-prolyl cis-trans isomerase activity	6.00	5	7.16
Ce	peptidyl-prolyl cis-trans isomerase activity	8.18	5	7.16
Dr	peptidyl-prolyl cis-trans isomerase activity	8.10	5	7.16
Ec	peptidyl-prolyl cis-trans isomerase activity		5	7.16
Pf	peptidyl-prolyl cis-trans isomerase activity	6.87	5	7.16
Tb	peptidyl-prolyl cis-trans isomerase activity	2.01	5	7.16
Hs	cysteine-type deubiquitinase activity	10.51	6	7.29
Mm	cysteine-type deubiquitinase activity	7.83	6	7.29
Sc	cysteine-type deubiquitinase activity		6	7.29
Dm	cysteine-type deubiquitinase activity	5.32	6	7.29
At	cysteine-type deubiquitinase activity	4.25	6	7.29
Ce	cysteine-type deubiquitinase activity	8.85	6	7.29
Dr	cysteine-type deubiquitinase activity	4.44	6	7.29
Ec	cysteine-type deubiquitinase activity		6	7.29
Pf	cysteine-type deubiquitinase activity	11.90	6	7.29
Tb	cysteine-type deubiquitinase activity	5.23	6	7.29
Hs	ATP-dependent activity, acting on DNA	5.72	9	7.52
Mm	ATP-dependent activity, acting on DNA	9.72	9	7.52
Sc	ATP-dependent activity, acting on DNA	10.28	9	7.52
Dm	ATP-dependent activity, acting on DNA	5.09	9	7.52
At	ATP-dependent activity, acting on DNA	6.21	9	7.52
Ce	ATP-dependent activity, acting on DNA	2.27	9	7.52
Dr	ATP-dependent activity, acting on DNA	7.67	9	7.52
Ec	ATP-dependent activity, acting on DNA	17.04	9	7.52
Pf	ATP-dependent activity, acting on DNA	6.87	9	7.52
Tb	ATP-dependent activity, acting on DNA	4.29	9	7.52
Hs	catalytic activity, acting on a tRNA		7	7.53
Mm	catalytic activity, acting on a tRNA	9.03	7	7.53
Sc	catalytic activity, acting on a tRNA	3.37	7	7.53
Dm	catalytic activity, acting on a tRNA	6.10	7	7.53
At	catalytic activity, acting on a tRNA	6.17	7	7.53
Ce	catalytic activity, acting on a tRNA	10.84	7	7.53
Dr	catalytic activity, acting on a tRNA	8.27	7	7.53
Ec	catalytic activity, acting on a tRNA		7	7.53
Pf	catalytic activity, acting on a tRNA		7	7.53
Tb	catalytic activity, acting on a tRNA	8.91	7	7.53
Hs	ribonuclease activity		6	7.61
Mm	ribonuclease activity	11.43	6	7.61
Sc	ribonuclease activity	2.29	6	7.61
Dm	ribonuclease activity	5.77	6	7.61
At	ribonuclease activity	7.42	6	7.61
Ce	ribonuclease activity	10.49	6	7.61
Dr	ribonuclease activity	9.01	6	7.61

Ec	ribonuclease activity		6	7.61
Pf	ribonuclease activity		6	7.61
Tb	ribonuclease activity	6.85	6	7.61
Hs	oxidoreductase activity, acting on a sulfur group of donors	11.35	6	7.63
Mm	oxidoreductase activity, acting on a sulfur group of donors	8.56	6	7.63
Sc	oxidoreductase activity, acting on a sulfur group of donors		6	7.63
Dm	oxidoreductase activity, acting on a sulfur group of donors	8.87	6	7.63
At	oxidoreductase activity, acting on a sulfur group of donors	3.38	6	7.63
Ce	oxidoreductase activity, acting on a sulfur group of donors	11.48	6	7.63
Dr	oxidoreductase activity, acting on a sulfur group of donors	4.45	6	7.63
Ec	oxidoreductase activity, acting on a sulfur group of donors		6	7.63
Pf	oxidoreductase activity, acting on a sulfur group of donors	10.30	6	7.63
Tb	oxidoreductase activity, acting on a sulfur group of donors	2.64	6	7.63
Hs	intramolecular oxidoreductase activity	7.61	7	7.66
Mm	intramolecular oxidoreductase activity	4.49	7	7.66
Sc	intramolecular oxidoreductase activity	4.39	7	7.66
Dm	intramolecular oxidoreductase activity	10.05	7	7.66
At	intramolecular oxidoreductase activity	3.95	7	7.66
Ce	intramolecular oxidoreductase activity	9.35	7	7.66
Dr	intramolecular oxidoreductase activity	4.95	7	7.66
Ec	intramolecular oxidoreductase activity		7	7.66
Pf	intramolecular oxidoreductase activity	17.86	7	7.66
Tb	intramolecular oxidoreductase activity	6.28	7	7.66
Hs	helicase activity	4.41	6	7.91
Mm	helicase activity	7.89	6	7.91
Sc	helicase activity	9.99	6	7.91
Dm	helicase activity	3.05	6	7.91
At	helicase activity	4.71	6	7.91
Ce	helicase activity	3.12	6	7.91
Dr	helicase activity	10.81	6	7.91
Ec	helicase activity	21.91	6	7.91
Pf	helicase activity		6	7.91
Tb	helicase activity	5.29	6	7.91
Hs	endoribonuclease activity		6	8.09
Mm	endoribonuclease activity	8.41	6	8.09
Sc	endoribonuclease activity	2.78	6	8.09
Dm	endoribonuclease activity	5.79	6	8.09
At	endoribonuclease activity	10.59	6	8.09
Ce	endoribonuclease activity	13.17	6	8.09
Dr	endoribonuclease activity	8.85	6	8.09
Ec	endoribonuclease activity		6	8.09
Pf	endoribonuclease activity		6	8.09
Tb	endoribonuclease activity	7.07	6	8.09

Hs	3'-5' exonuclease activity	7.07	5	8.15
Mm	3'-5' exonuclease activity	9.86	5	8.15
Sc	3'-5' exonuclease activity		5	8.15
Dm	3'-5' exonuclease activity	5.32	5	8.15
At	3'-5' exonuclease activity	5.85	5	8.15
Ce	3'-5' exonuclease activity	3.84	5	8.15
Dr	3'-5' exonuclease activity	8.58	5	8.15
Ec	3'-5' exonuclease activity	13.28	5	8.15
Pf	3'-5' exonuclease activity	13.95	5	8.15
Tb	3'-5' exonuclease activity	5.65	5	8.15
Hs	histone binding	11.64	7	8.20
Mm	histone binding	6.71	7	8.20
Sc	histone binding	8.30	7	8.20
Dm	histone binding	9.59	7	8.20
At	histone binding	6.78	7	8.20
Ce	histone binding	11.16	7	8.20
Dr	histone binding	6.08	7	8.20
Ec	histone binding		7	8.20
Pf	histone binding	5.36	7	8.20
Tb	histone binding		7	8.20
Hs	modification-dependent protein binding	9.26	7	8.47
Mm	modification-dependent protein binding	6.29	7	8.47
Sc	modification-dependent protein binding	13.91	7	8.47
Dm	modification-dependent protein binding	11.07	7	8.47
At	modification-dependent protein binding	3.67	7	8.47
Ce	modification-dependent protein binding	12.03	7	8.47
Dr	modification-dependent protein binding	5.97	7	8.47
Ec	modification-dependent protein binding		7	8.47
Pf	modification-dependent protein binding	5.58	7	8.47
Tb	modification-dependent protein binding		7	8.47
Hs	oxidoreductase activity, acting on the aldehyde or oxo group of donors	14.74	7	8.94
Mm	oxidoreductase activity, acting on the aldehyde or oxo group of donors	8.66	7	8.94
Sc	oxidoreductase activity, acting on the aldehyde or oxo group of donors		7	8.94
Dm	oxidoreductase activity, acting on the aldehyde or oxo group of donors	6.03	7	8.94
At	oxidoreductase activity, acting on the aldehyde or oxo group of donors	3.63	7	8.94
Ce	oxidoreductase activity, acting on the aldehyde or oxo group of donors	12.59	7	8.94
Dr	oxidoreductase activity, acting on the aldehyde or oxo group of donors	8.59	7	8.94
Ec	oxidoreductase activity, acting on the aldehyde or oxo group of donors	13.46	7	8.94
Pf	oxidoreductase activity, acting on the aldehyde or oxo group of donors		7	8.94
Tb	oxidoreductase activity, acting on the aldehyde or oxo group of donors	3.86	7	8.94
Hs	disulfide oxidoreductase activity	15.39	6	9.02
Mm	disulfide oxidoreductase activity	11.04	6	9.02
Sc	disulfide oxidoreductase activity	4.12	6	9.02

Dm	disulfide oxidoreductase activity	11.75	6	9.02
At	disulfide oxidoreductase activity	2.54	6	9.02
Ce	disulfide oxidoreductase activity	15.22	6	9.02
Dr	disulfide oxidoreductase activity	6.07	6	9.02
Ec	disulfide oxidoreductase activity	10.96	6	9.02
Pf	disulfide oxidoreductase activity	8.93	6	9.02
Tb	disulfide oxidoreductase activity	4.19	6	9.02
Hs	RNA methyltransferase activity		5	9.09
Mm	RNA methyltransferase activity	8.28	5	9.09
Sc	RNA methyltransferase activity		5	9.09
Dm	RNA methyltransferase activity		5	9.09
At	RNA methyltransferase activity	5.03	5	9.09
Ce	RNA methyltransferase activity	15.59	5	9.09
Dr	RNA methyltransferase activity	8.70	5	9.09
Ec	RNA methyltransferase activity		5	9.09
Pf	RNA methyltransferase activity		5	9.09
Tb	RNA methyltransferase activity	7.85	5	9.09
Hs	ligase activity, forming carbon-oxygen bonds		6	9.18
Mm	ligase activity, forming carbon-oxygen bonds	9.69	6	9.18
Sc	ligase activity, forming carbon-oxygen bonds		6	9.18
Dm	ligase activity, forming carbon-oxygen bonds	6.84	6	9.18
At	ligase activity, forming carbon-oxygen bonds	7.63	6	9.18
Ce	ligase activity, forming carbon-oxygen bonds	6.63	6	9.18
Dr	ligase activity, forming carbon-oxygen bonds	12.87	6	9.18
Ec	ligase activity, forming carbon-oxygen bonds		6	9.18
Pf	ligase activity, forming carbon-oxygen bonds		6	9.18
Tb	ligase activity, forming carbon-oxygen bonds	11.42	6	9.18
Hs	aminoacyl-tRNA ligase activity		6	9.18
Mm	aminoacyl-tRNA ligase activity	9.69	6	9.18
Sc	aminoacyl-tRNA ligase activity		6	9.18
Dm	aminoacyl-tRNA ligase activity	6.84	6	9.18
At	aminoacyl-tRNA ligase activity	7.63	6	9.18
Ce	aminoacyl-tRNA ligase activity	6.63	6	9.18
Dr	aminoacyl-tRNA ligase activity	12.87	6	9.18
Ec	aminoacyl-tRNA ligase activity		6	9.18
Pf	aminoacyl-tRNA ligase activity		6	9.18
Tb	aminoacyl-tRNA ligase activity	11.42	6	9.18
Hs	5'-3' RNA polymerase activity		6	9.22
Mm	5'-3' RNA polymerase activity	13.25	6	9.22
Sc	5'-3' RNA polymerase activity	5.22	6	9.22
Dm	5'-3' RNA polymerase activity	12.66	6	9.22
At	5'-3' RNA polymerase activity	6.69	6	9.22
Ce	5'-3' RNA polymerase activity	12.40	6	9.22
Dr	5'-3' RNA polymerase activity	9.45	6	9.22
Ec	5'-3' RNA polymerase activity		6	9.22
Pf	5'-3' RNA polymerase activity	5.58	6	9.22
Tb	5'-3' RNA polymerase activity	8.56	6	9.22
Hs	RNA polymerase activity		6	9.22
Mm	RNA polymerase activity	13.25	6	9.22
Sc	RNA polymerase activity	5.22	6	9.22
Dm	RNA polymerase activity	12.66	6	9.22
At	RNA polymerase activity	6.69	6	9.22
Ce	RNA polymerase activity	12.40	6	9.22

Dr	RNA polymerase activity	9.45	6	9.22
Ec	RNA polymerase activity		6	9.22
Pf	RNA polymerase activity	5.58	6	9.22
Tb	RNA polymerase activity	8.56	6	9.22
Hs	DNA helicase activity	6.69	5	9.26
Mm	DNA helicase activity	9.46	5	9.26
Sc	DNA helicase activity	14.17	5	9.26
Dm	DNA helicase activity	3.50	5	9.26
At	DNA helicase activity	6.86	5	9.26
Ce	DNA helicase activity		5	9.26
Dr	DNA helicase activity	8.52	5	9.26
Ec	DNA helicase activity	21.48	5	9.26
Pf	DNA helicase activity		5	9.26
Tb	DNA helicase activity	3.43	5	9.26
Hs	nucleosome binding	5.74	5	9.36
Mm	nucleosome binding	6.48	5	9.36
Sc	nucleosome binding	5.06	5	9.36
Dm	nucleosome binding	14.15	5	9.36
At	nucleosome binding	10.75	5	9.36
Ce	nucleosome binding	12.27	5	9.36
Dr	nucleosome binding	8.66	5	9.36
Ec	nucleosome binding		5	9.36
Pf	nucleosome binding	14.88	5	9.36
Tb	nucleosome binding	6.28	5	9.36
Hs	DNA-directed 5'-3' RNA polymerase activity		6	9.57
Mm	DNA-directed 5'-3' RNA polymerase activity	13.04	6	9.57
Sc	DNA-directed 5'-3' RNA polymerase activity	5.56	6	9.57
Dm	DNA-directed 5'-3' RNA polymerase activity	12.66	6	9.57
At	DNA-directed 5'-3' RNA polymerase activity	7.73	6	9.57
Ce	DNA-directed 5'-3' RNA polymerase activity	14.11	6	9.57
Dr	DNA-directed 5'-3' RNA polymerase activity	9.28	6	9.57
Ec	DNA-directed 5'-3' RNA polymerase activity		6	9.57
Pf	DNA-directed 5'-3' RNA polymerase activity	5.58	6	9.57
Tb	DNA-directed 5'-3' RNA polymerase activity	8.56	6	9.57
Hs	protein-disulfide reductase activity	14.84	6	9.77
Mm	protein-disulfide reductase activity	11.33	6	9.77
Sc	protein-disulfide reductase activity		6	9.77
Dm	protein-disulfide reductase activity	11.67	6	9.77
At	protein-disulfide reductase activity	2.73	6	9.77
Ce	protein-disulfide reductase activity	16.59	6	9.77
Dr	protein-disulfide reductase activity	5.47	6	9.77
Ec	protein-disulfide reductase activity		6	9.77
Pf	protein-disulfide reductase activity	9.92	6	9.77
Tb	protein-disulfide reductase activity	5.58	6	9.77
Hs	tRNA methyltransferase activity		5	9.84
Mm	tRNA methyltransferase activity	10.35	5	9.84
Sc	tRNA methyltransferase activity		5	9.84
Dm	tRNA methyltransferase activity		5	9.84
At	tRNA methyltransferase activity	6.33	5	9.84
Ce	tRNA methyltransferase activity	15.07	5	9.84
Dr	tRNA methyltransferase activity	8.48	5	9.84
Ec	tRNA methyltransferase activity		5	9.84
Pf	tRNA methyltransferase activity		5	9.84

Tb	tRNA methyltransferase activity	8.97	5	9.84
Hs	peptide-lysine-N-acetyltransferase activity	10.99	5	10.36
Mm	peptide-lysine-N-acetyltransferase activity	6.97	5	10.36
Sc	peptide-lysine-N-acetyltransferase activity		5	10.36
Dm	peptide-lysine-N-acetyltransferase activity		5	10.36
At	peptide-lysine-N-acetyltransferase activity	6.47	5	10.36
Ce	peptide-lysine-N-acetyltransferase activity	21.43	5	10.36
Dr	peptide-lysine-N-acetyltransferase activity	5.94	5	10.36
Ec	peptide-lysine-N-acetyltransferase activity		5	10.36
Pf	peptide-lysine-N-acetyltransferase activity		5	10.36
Tb	peptide-lysine-N-acetyltransferase activity		5	10.36
Hs	histone acetyltransferase activity	11.99	5	10.69
Mm	histone acetyltransferase activity	7.57	5	10.69
Sc	histone acetyltransferase activity		5	10.69
Dm	histone acetyltransferase activity	6.03	5	10.69
At	histone acetyltransferase activity	6.78	5	10.69
Ce	histone acetyltransferase activity	25.00	5	10.69
Dr	histone acetyltransferase activity	6.75	5	10.69
Ec	histone acetyltransferase activity		5	10.69
Pf	histone acetyltransferase activity		5	10.69
Tb	histone acetyltransferase activity		5	10.69
Hs	oxidoreductase activity, acting on the aldehyde or oxo group of donors, NAD or NADP as acceptor	17.76	7	11.55
Sc	RNA 3'-end processing		6	8.03
Dm	RNA 3'-end processing	6.75	6	8.03
At	RNA 3'-end processing	5.27	6	8.03
Ce	RNA 3'-end processing	13.09	6	8.03
Dr	RNA 3'-end processing	9.20	6	8.03
Ec	RNA 3'-end processing		6	8.03
Pf	RNA 3'-end processing		6	8.03
Tb	RNA 3'-end processing	6.65	6	8.03
Hs	RNA methylation		5	8.03
Mm	RNA methylation	8.51	5	8.03
Sc	RNA methylation	3.39	5	8.03
Dm	RNA methylation		5	8.03
At	RNA methylation	4.74	5	8.03
Ce	RNA methylation	15.55	5	8.03
Dr	RNA methylation	8.16	5	8.03
Ec	RNA methylation		5	8.03
Pf	RNA methylation		5	8.03
Tb	RNA methylation	7.81	5	8.03
Hs	demethylation	10.82	5	8.10
Mm	demethylation	7.41	5	8.10
Sc	demethylation	9.27	5	8.10
Dm	demethylation		5	8.10
At	demethylation	4.35	5	8.10
Ce	demethylation	10.23	5	8.10
Dr	demethylation	6.25	5	8.10
Ec	demethylation		5	8.10
Pf	demethylation		5	8.10
Tb	demethylation	8.37	5	8.10
Hs	amino acid activation		6	8.12

Mm	amino acid activation	9.78	6	8.12
Sc	amino acid activation	3.48	6	8.12
Dm	amino acid activation	6.89	6	8.12
At	amino acid activation	7.91	6	8.12
Ce	amino acid activation	5.84	6	8.12
Dr	amino acid activation	11.88	6	8.12
Ec	amino acid activation		6	8.12
Pf	amino acid activation		6	8.12
Tb	amino acid activation	11.05	6	8.12
Hs	DNA geometric change	6.06	5	8.15
Mm	DNA geometric change	5.41	5	8.15
Sc	DNA geometric change	13.41	5	8.15
Dm	DNA geometric change	4.02	5	8.15
At	DNA geometric change	7.19	5	8.15
Ce	DNA geometric change		5	8.15
Dr	DNA geometric change	5.12	5	8.15
Ec	DNA geometric change	21.48	5	8.15
Pf	DNA geometric change		5	8.15
Tb	DNA geometric change	2.51	5	8.15
Hs	DNA duplex unwinding	6.45	5	8.25
Mm	DNA duplex unwinding	5.52	5	8.25
Sc	DNA duplex unwinding	13.41	5	8.25
Dm	DNA duplex unwinding	4.34	5	8.25
At	DNA duplex unwinding	7.19	5	8.25
Ce	DNA duplex unwinding		5	8.25
Dr	DNA duplex unwinding	5.12	5	8.25
Ec	DNA duplex unwinding	21.48	5	8.25
Pf	DNA duplex unwinding		5	8.25
Tb	DNA duplex unwinding	2.51	5	8.25
Hs	tRNA aminoacylation		6	8.28
Mm	tRNA aminoacylation	10.01	6	8.28
Sc	tRNA aminoacylation	3.48	6	8.28
Dm	tRNA aminoacylation	7.23	6	8.28
At	tRNA aminoacylation	7.91	6	8.28
Ce	tRNA aminoacylation	6.14	6	8.28
Dr	tRNA aminoacylation	12.12	6	8.28
Ec	tRNA aminoacylation		6	8.28
Pf	tRNA aminoacylation		6	8.28
Tb	tRNA aminoacylation	11.05	6	8.28
Hs	tRNA aminoacylation for protein translation		6	8.31
Mm	tRNA aminoacylation for protein translation	9.93	6	8.31
Sc	tRNA aminoacylation for protein translation	3.01	6	8.31
Dm	tRNA aminoacylation for protein translation	6.84	6	8.31
At	tRNA aminoacylation for protein translation	7.91	6	8.31
Ce	tRNA aminoacylation for protein translation	6.63	6	8.31
Dr	tRNA aminoacylation for protein translation	12.82	6	8.31
Ec	tRNA aminoacylation for protein translation		6	8.31
Pf	tRNA aminoacylation for protein translation		6	8.31
Tb	tRNA aminoacylation for protein translation	10.99	6	8.31
Hs	regulation of DNA recombination	5.33	5	8.39
Mm	regulation of DNA recombination	4.88	5	8.39
Sc	regulation of DNA recombination	5.06	5	8.39
Dm	regulation of DNA recombination	15.23	5	8.39

At	regulation of DNA recombination	7.91	5	8.39
Ce	regulation of DNA recombination	12.03	5	8.39
Dr	regulation of DNA recombination	8.29	5	8.39
Ec	regulation of DNA recombination		5	8.39
Pf	regulation of DNA recombination		5	8.39
Tb	regulation of DNA recombination		5	8.39
Hs	mitochondrial RNA metabolic process		5	9.28
Mm	mitochondrial RNA metabolic process	5.65	5	9.28
Sc	mitochondrial RNA metabolic process	6.55	5	9.28
Dm	mitochondrial RNA metabolic process		5	9.28
At	mitochondrial RNA metabolic process	8.36	5	9.28
Ce	mitochondrial RNA metabolic process	16.36	5	9.28
Dr	mitochondrial RNA metabolic process	9.50	5	9.28
Ec	mitochondrial RNA metabolic process		5	9.28
Pf	mitochondrial RNA metabolic process		5	9.28
Tb	mitochondrial RNA metabolic process		5	9.28
Hs	negative regulation of DNA metabolic process	4.68	5	9.29
Mm	negative regulation of DNA metabolic process	4.27	5	9.29
Sc	negative regulation of DNA metabolic process	3.57	5	9.29
Dm	negative regulation of DNA metabolic process	9.04	5	9.29
At	negative regulation of DNA metabolic process	6.89	5	9.29
Ce	negative regulation of DNA metabolic process	11.16	5	9.29
Dr	negative regulation of DNA metabolic process	7.97	5	9.29
Ec	negative regulation of DNA metabolic process		5	9.29
Pf	negative regulation of DNA metabolic process	29.76	5	9.29
Tb	negative regulation of DNA metabolic process	6.28	5	9.29
Hs	ncRNA catabolic process		5	9.77
Mm	ncRNA catabolic process	8.28	5	9.77
Sc	ncRNA catabolic process		5	9.77
Dm	ncRNA catabolic process	12.05	5	9.77
At	ncRNA catabolic process	4.61	5	9.77
Ce	ncRNA catabolic process		5	9.77
Dr	ncRNA catabolic process	12.50	5	9.77
Ec	ncRNA catabolic process		5	9.77
Pf	ncRNA catabolic process		5	9.77
Tb	ncRNA catabolic process	11.42	5	9.77
Hs	RNA polyadenylation		5	10.09
Mm	RNA polyadenylation	7.24	5	10.09
Sc	RNA polyadenylation		5	10.09
Dm	RNA polyadenylation	8.68	5	10.09
At	RNA polyadenylation	4.87	5	10.09
Ce	RNA polyadenylation	24.55	5	10.09
Dr	RNA polyadenylation	10.39	5	10.09
Ec	RNA polyadenylation		5	10.09
Pf	RNA polyadenylation		5	10.09
Tb	RNA polyadenylation	4.83	5	10.09
Hs	negative regulation of DNA recombination	6.60	5	11.17
Mm	negative regulation of DNA recombination	6.34	5	11.17
Sc	negative regulation of DNA recombination		5	11.17
Dm	negative regulation of DNA recombination	18.08	5	11.17
At	negative regulation of DNA recombination	9.16	5	11.17
Ce	negative regulation of DNA recombination	15.74	5	11.17
Dr	negative regulation of DNA recombination	11.13	5	11.17

Ec	negative regulation of DNA recombination		5	11.17
Pf	negative regulation of DNA recombination		5	11.17
Tb	negative regulation of DNA recombination		5	11.17
Hs	cell redox homeostasis	7.07	5	11.23
Mm	cell redox homeostasis		5	11.23
Sc	cell redox homeostasis		5	11.23
Dm	cell redox homeostasis	12.05	5	11.23
At	cell redox homeostasis	6.68	5	11.23
Ce	cell redox homeostasis	12.59	5	11.23
Dr	cell redox homeostasis	7.03	5	11.23
Ec	cell redox homeostasis		5	11.23
Pf	cell redox homeostasis	26.04	5	11.23
Tb	cell redox homeostasis	7.18	5	11.23
Hs	DNA-templated DNA replication maintenance of fidelity		6	14.89
Mm	DNA-templated DNA replication maintenance of fidelity	5.19	6	14.89
Sc	DNA-templated DNA replication maintenance of fidelity		6	14.89
Dm	DNA-templated DNA replication maintenance of fidelity		6	14.89
At	DNA-templated DNA replication maintenance of fidelity	7.38	6	14.89
Ce	DNA-templated DNA replication maintenance of fidelity	11.16	6	14.89
Dr	DNA-templated DNA replication maintenance of fidelity	5.14	6	14.89
Ec	DNA-templated DNA replication maintenance of fidelity	31.30	6	14.89
Pf	DNA-templated DNA replication maintenance of fidelity	35.71	6	14.89
Tb	DNA-templated DNA replication maintenance of fidelity	8.37	6	14.89
Hs	tRNA-type intron splice site recognition and cleavage		5	28.11
Mm	tRNA-type intron splice site recognition and cleavage	33.11	5	28.11
Sc	tRNA-type intron splice site recognition and cleavage	27.82	5	28.11
Dm	tRNA-type intron splice site recognition and cleavage	36.16	5	28.11
At	tRNA-type intron splice site recognition and cleavage	15.81	5	28.11
Ce	tRNA-type intron splice site recognition and cleavage	40.91	5	28.11
Dr	tRNA-type intron splice site recognition and cleavage	14.84	5	28.11
Ec	tRNA-type intron splice site recognition and cleavage		5	28.11
Pf	tRNA-type intron splice site recognition and cleavage		5	28.11
Tb	tRNA-type intron splice site recognition and cleavage		5	28.11

Table 11: List of newly discovered RBDs and their enrichment

InterPro ID	InterPro Name	Reference dataset	Fold enrichment	Adjusted p-value
IPR000225	Armadillo	non_RBP_no_RBD	5.68	2.36E-02
IPR001680	WD40 repeat	low_score	7.40	3.24E-07
IPR001680	WD40 repeat	non_RBP_no_RBD	5.90	1.44E-16
IPR001680	WD40 repeat	proteome	4.07	6.45E-12
IPR001715	Calponin homology domain	non_RBP_no_RBD	6.37	5.31E-04
IPR001715	Calponin homology domain	proteome	3.24	2.36E-02
IPR002017	Spectrin repeat	non_RBP_no_RBD	13.69	4.47E-04
IPR002017	Spectrin repeat	proteome	7.71	2.91E-03
IPR002125	Cytidine and deoxycytidylate deaminase domain	non_RBP_no_RBD	38.51	1.57E-03
IPR002125	Cytidine and deoxycytidylate deaminase domain	proteome	6.93	4.27E-02
IPR002130	Cyclophilin-type peptidyl-prolyl cis-trans isomerase domain	non_RBP_no_RBD	25.67	1.28E-06
IPR002130	Cyclophilin-type peptidyl-prolyl cis-trans isomerase domain	proteome	9.24	2.41E-04
IPR002652	Importin-alpha, importin-beta-binding domain	non_RBP_no_RBD	25.67	2.95E-03
IPR002652	Importin-alpha, importin-beta-binding domain	proteome	12.32	8.65E-03
IPR003594	Histidine kinase/HSP90-like ATPase	non_RBP_no_RBD	18.49	1.62E-03
	Histidine kinase/HSP90-like ATPase	proteome	8.87	8.65E-03
IPR003959	ATPase, AAA-type, core	non_RBP_no_RBD	6.85	6.31E-03
IPR007052	CS domain	non_RBP_no_RBD	19.26	5.31E-03
IPR007052	CS domain	proteome	6.93	4.27E-02
IPR018159	Spectrin/alpha-actinin	low_score	Inf	3.85E-02
IPR018159	Spectrin/alpha-actinin	non_RBP_no_RBD	10.66	4.47E-04
IPR018159	Spectrin/alpha-actinin	proteome	6.65	2.62E-03
IPR018502	Annexin repeat	non_RBP_no_RBD	12.84	1.06E-02
IPR018502	Annexin repeat	proteome	7.92	2.96E-02
IPR019734	Tetratricopeptide repeat	non_RBP_no_RBD	4.36	4.47E-04
IPR019734	Tetratricopeptide repeat	proteome	2.77	1.25E-02
IPR020472	G-protein beta WD-40 repeat	non_RBP_no_RBD	9.84	1.02E-10

IPR020472	G-protein beta WD-40 repeat	low_score	8.10	2.56E-03
IPR020472	G-protein beta WD-40 repeat	proteome	6.00	8.54E-08
IPR032413	Atypical Arm repeat	non_RBP_no_RBD	38.51	1.57E-03
IPR032413	Atypical Arm repeat	proteome	15.83	5.50E-03

8. References

1. Crick,F. (1970) Central Dogma of Molecular Biology. *Nature*, **227**, 561–563.
2. Crick, F.H.C. (1958) On protein synthesis. *Symp. Soc. Exp. Biol.*, **12**, 138–63.
3. Gall,J.G. (1956) SMALL GRANULES IN THE AMPHIBIAN OOCYTE NUCLEUS AND THEIR RELATIONSHIP TO RNA. *The Journal of Biophysical and Biochemical Cytology*, **2**, 393–396.
4. Swift,H. (1963) Cytochemical studies on nuclear fine structure. *Experimental Cell Research*, **9**, 54–67.
5. Malcolm,D.B. and Sommerville,J. (1974) The structure of chromosome-derived ribonucleoprotein in oocytes of *Triturus cristatus carnifex* (Laurenti). *Chromosoma*, **48**, 137–158.
6. Dreyfuss,G., Matunis,M.J., Pinol-Roma,S. and Burd,C.G. (1993) hnRNP PROTEINS AND THE BIOGENESIS OF mRNA. *Annual Review of Biochemistry*, **62**, 289–321.
7. Dreyfuss,G., Kim,V.N. and Kataoka,N. (2002) Messenger-RNA-binding proteins and the messages they carry. *Nature Reviews Molecular Cell Biology*, **3**, 195–205.
8. Spirin,A.S., Belitsina,N.V. and Lerman,M.I. (1965) Use of formaldehyde fixation for studies of ribonucleoprotein particles by caesium chloride density-gradient centrifugation. *Journal of Molecular Biology*, **14**, 611-IN30.
9. Dreyfuss,G. (1986) Structure and Function of Nuclear and Cytoplasmic Ribonucleoprotein Particles.
10. Dreyfuss,G., Swanson,M.S. and Piñol-Roma,S. (1988) Heterogeneous nuclear ribonucleoprotein particles and the pathway of mRNA formation. *Trends in Biochemical Sciences*, **13**, 86–91.
11. Michelotti,E.F., Michelotti,G.A., Aronsohn,A.I. and Levens,D. (1996) Heterogeneous nuclear ribonucleoprotein K is a transcription factor. *Molecular and Cellular Biology*, **16**, 2350–2360.
12. Miao,L.-H., Chang,C.-J., Shen,B.-J., Tsai,W.-H. and Lee,S.-C. (1998) Identification of Heterogeneous Nuclear Ribonucleoprotein K (hnRNP K) as a Repressor of C/EBP β -mediated Gene Activation *. *Journal of Biological Chemistry*, **273**, 10784–10791.
13. Mayeda,A. and Krainer,F. (1992) Regulation of Alternative Pre-mRNA Splicing by hnRNP A1 and Splicing Factor SF2.
14. Cáceres,J.F., Stamm,S., Helfman,D.M. and Krainer,A.R. (1994) Regulation of Alternative Splicing in Vivo by Overexpression of Antagonistic Splicing Factors. *Science*, **265**, 1706–1709.
15. Bagga,P.S., Arhin,G.K. and Wilusz,J. (1998) DSEF-1 is a member of the hnRNP H family of RNA-binding proteins and stimulates pre-mRNA cleavage and polyadenylation in vitro. *Nucleic Acids Research*, **26**, 5343–5350.
16. Pollard,V.W., Michael,W.M., Nakielny,S., Siomi,M.C., Wang,F. and Dreyfuss,G. (1996) A Novel Receptor-Mediated Nuclear Protein Import Pathway. *Cell*, **86**, 985–994.

17. Izaurrealde,E., Jarmolowski,A., Beisel,C., Mattaj,I.W., Dreyfuss,G. and Fischer,U. (1997) A Role for the M9 Transport Signal of hnRNP A1 in mRNA Nuclear Export. *Journal of Cell Biology*, **137**, 27–35.
18. Ostareck,D.H., Ostareck-Lederer,A., Shatsky,I.N. and Hentze,M.W. (2001) Lipoygenase mRNA Silencing in Erythroid Differentiation: The 3'UTR Regulatory Complex Controls 60S Ribosomal Subunit Joining. *Cell*, **104**, 281–290.
19. Habelhah,H., Shah,K., Huang,L., Ostareck-Lederer,A., Burlingame,A.L., Shokat,K.M., Hentze,M.W. and Ronai,Z. (2001) ERK phosphorylation drives cytoplasmic accumulation of hnRNP-K and inhibition of mRNA translation. *Nat Cell Biol*, **3**, 325–330.
20. Hir,H.L., Moore,M.J. and Maquat,L.E. (2000) Pre-mRNA splicing alters mRNP composition: evidence for stable association of proteins at exon–exon junctions. *Genes Dev.*, **14**, 1098–1108.
21. Luo,M. and Reed,R. (1999) Splicing is required for rapid and efficient mRNA export in metazoans. *Proceedings of the National Academy of Sciences*, **96**, 14937–14942.
22. Zhou,Z., Luo,M., Straesser,K., Katahira,J., Hurt,E. and Reed,R. (2000) The protein Aly links pre-messenger-RNA splicing to nuclear export in metazoans. *Nature*, **407**, 401–405.
23. Sun,X., Moriarty,P.M. and Maquat,L.E. (2000) Nonsense-mediated decay of glutathione peroxidase 1 mRNA in the cytoplasm depends on intron position. *The EMBO Journal*, **19**, 4734–4744.
24. Hentze,M.W., Castello,A., Schwarzl,T. and Preiss,T. (2018) A brave new world of RNA-binding proteins. *Nature Reviews Molecular Cell Biology*, 10.1038/nrm.2017.130.
25. Rinn,J.L. and Chang,H.Y. (2012) Genome Regulation by Long Noncoding RNAs. *Annual Review of Biochemistry*, **81**, 145–166.
26. Margueron,R. and Reinberg,D. (2011) The Polycomb complex PRC2 and its mark in life. *Nature*, **469**, 343–349.
27. Shi,Y., Lan,F., Matson,C., Mulligan,P., Whetstine,J.R., Cole,P.A., Casero,R.A. and Shi,Y. (2004) Histone Demethylation Mediated by the Nuclear Amine Oxidase Homolog LSD1. *Cell*, **119**, 941–953.
28. Lunyak,V.V., Burgess,R., Prefontaine,G.G., Nelson,C., Sze,S.-H., Chenoweth,J., Schwartz,P., Pevzner,P.A., Glass,C., Mandel,G., *et al.* (2002) Corepressor-dependent silencing of chromosomal regions encoding neuronal genes. *Science*, **298**, 1747–1752.
29. Tsai,M.-C., Manor,O., Wan,Y., Mosammamaparast,N., Wang,J.K., Lan,F., Shi,Y., Segal,E. and Chang,H.Y. (2010) Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes. *Science*, **329**, 689–693.
30. Kino,T., Hurt,D.E., Ichijo,T., Nader,N. and Chrousos,G.P. (2010) Noncoding RNA Gas5 Is a Growth Arrest– and Starvation-Associated Repressor of the Glucocorticoid Receptor. *Sci. Signal.*, **3**, ra8–ra8.
31. Alberti,S. (2017) Phase separation in biology. *Current Biology*, **27**, R1097–R1102.
32. Schwartz,J.C., Wang,X., Podell,E.R. and Cech,T.R. (2013) RNA Seeds Higher-Order Assembly of FUS Protein. *Cell Reports*, **5**, 918–925.

33. Burke,K.A., Janke,A.M., Rhine,C.L. and Fawzi,N.L. (2015) Residue-by-Residue View of In Vitro FUS Granules that Bind the C-Terminal Domain of RNA Polymerase II. *Molecular Cell*, **60**, 231–241.
34. Maharana,S., Wang,J., Papadopoulos,D.K., Richter,D., Pozniakovsky,A., Poser,I., Bickle,M., Rizk,S., Guillén-Boixet,J., Franzmann,T.M., *et al.* (2018) RNA buffers the phase separation behavior of prion-like RNA binding proteins. *Science*, **360**, 918–921.
35. Elbaum-Garfinkle,S., Kim,Y., Szczepaniak,K., Chen,C.C.-H., Eckmann,C.R., Myong,S. and Brangwynne,C.P. (2015) The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics. *Proceedings of the National Academy of Sciences*, **112**, 7189–7194.
36. Pouloupoulos,A., Murphy,A.J., Ozkan,A., Davis,P., Hatch,J., Kirchner,R. and Macklis,J.D. (2019) Subcellular transcriptomes and proteomes of developing axon projections in the cerebral cortex. *Nature*, **565**, 356–360.
37. Lukong,K.E., Chang,K., Khandjian,E.W. and Richard,S. (2008) RNA-binding proteins in human genetic disease. *Trends in Genetics*, **24**, 416–425.
38. Nussbacher,J.K., Tabet,R., Yeo,G.W. and Lagier-Tourenne,C. (2019) Disruption of RNA Metabolism in Neurological Diseases and Emerging Therapeutic Interventions. *Neuron*, **102**, 294–320.
39. Oberlé,I., Rousseau,F., Heitz,D., Kretz,C., Devys,D., Hanauer,A., Boué,J., Bertheas,M.F. and Mandel,J.L. (1991) Instability of a 550-Base Pair DNA Segment and Abnormal Methylation in Fragile X Syndrome. *Science*, **252**, 1097–1102.
40. Verkerk,A.J.M.H., Pieretti,M., Sutcliffe,J.S., Fu,Y.-H., Kuhl,D.P.A., Pizzuti,A., Reiner,O., Richards,S., Victoria,M.F., Zhang,F., *et al.* (1991) Identification of a gene (FMR-1) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell*, **65**, 905–914.
41. Kremer,E.J., Pritchard,M., Lynch,M., Yu,S., Holman,K., Baker,E., Warren,S.T., Schlessinger,D., Sutherland,G.R. and Richards,R.I. (1991) Mapping of DNA Instability at the Fragile X to a Trinucleotide Repeat Sequence P(CCG)*n*. *Science*, **252**, 1711–1714.
42. Santoro,M.R., Bray,S.M. and Warren,S.T. (2012) Molecular Mechanisms of Fragile X Syndrome: A Twenty-Year Perspective. *Annu. Rev. Pathol. Mech. Dis.*, **7**, 219–245.
43. Shaw,C.E., Al-Chalabi,A. and Leigh,N. (2001) Progress in the pathogenesis of amyotrophic lateral sclerosis. *Curr Neurol Neurosci Rep*, **1**, 69–76.
44. Sreedharan,J., Blair,I.P., Tripathi,V.B., Hu,X., Vance,C., Rogelj,B., Ackerley,S., Durnall,J.C., Williams,K.L., Buratti,E., *et al.* (2008) TDP-43 Mutations in Familial and Sporadic Amyotrophic Lateral Sclerosis. *Science*, **319**, 1668–1672.
45. Scotter,E.L., Chen,H.-J. and Shaw,C.E. (2015) TDP-43 Proteinopathy and ALS: Insights into Disease Mechanisms and Therapeutic Targets. *Neurotherapeutics*, **12**, 352–363.
46. Meola,G. (2000) Myotonic dystrophies. *Current Opinion in Neurology*, **13**, 519.
47. Brook,J.D., McCurrach,M.E., Harley,H.G., Buckler,A.J., Church,D., Aburatani,H., Hunter,K., Stanton,V.P., Thirion,J.-P., Hudson,T., *et al.* (1992) Molecular basis of myotonic dystrophy:

Expansion of a trinucleotide (CTG) repeat at the 3' end of a transcript encoding a protein kinase family member. *Cell*, **68**, 799–808.

48. Liquori, C.L., Ricker, K., Moseley, M.L., Jacobsen, J.F., Kress, W., Naylor, S.L., Day, J.W. and Ranum, L.P.W. (2001) Myotonic dystrophy type 2 caused by a CCTG expansion in intron I of ZNF9. *Science*, **293**, 864–867.
49. Mankodi, A., Logigian, E., Callahan, L., McClain, C., White, R., Henderson, D., Krym, M. and Thornton, C.A. (2000) Myotonic Dystrophy in Transgenic Mice Expressing an Expanded CUG Repeat. *Science*, **289**, 1769–1772.
50. Wang, E.T., Cody, N.A.L., Jog, S., Biancolella, M., Wang, T.T., Treacy, D.J., Luo, S., Schroth, G.P., Housman, D.E., Reddy, S., *et al.* (2012) Transcriptome-wide Regulation of Pre-mRNA Splicing and mRNA Localization by Muscleblind Proteins. *Cell*, **150**, 710–724.
51. Hanahan, D. and Weinberg, R.A. (2011) Hallmarks of Cancer: The Next Generation. *Cell*, **144**, 646–674.
52. Pereira, B., Billaud, M. and Almeida, R. (2017) RNA-Binding Proteins in Cancer: Old Players and New Actors. *Trends in Cancer*, **3**, 506–528.
53. Wang, W., Caldwell, M.C., Lin, S., Furneaux, H. and Gorospe, M. (2000) HuR regulates cyclin A and cyclin B1 mRNA stability during cell proliferation. *The EMBO Journal*, **19**, 2340–2350.
54. Lal, A., Mazan-Mamczarz, K., Kawai, T., Yang, X., Martindale, J.L. and Gorospe, M. (2004) Concurrent versus individual binding of HuR and AUF1 to common labile target mRNAs. *The EMBO Journal*, **23**, 3092–3102.
55. Guo, X. and Hartley, R.S. (2006) HuR Contributes to Cyclin E1 Dereglulation in MCF-7 Breast Cancer Cells. *Cancer Research*, **66**, 7948–7956.
56. Abdelmohsen, K., Lal, A., Kim, H.H. and Gorospe, M. (2007) Posttranscriptional Orchestration of an Anti-Apoptotic Program by HuR. *Cell Cycle*, **6**, 1288–1292.
57. Schultz, C.W., Preet, R., Dhir, T., Dixon, D.A. and Brody, J.R. (2020) Understanding and targeting the disease-related RNA binding protein human antigen R (HuR). *WIREs RNA*, **11**, e1581.
58. Miyata, Y., Watanabe, S., Sagara, Y., Mitsunari, K., Matsuo, T., Ohba, K. and Sakai, H. (2013) High Expression of HuR in Cytoplasm, but Not Nuclei, Is Associated with Malignant Aggressiveness and Prognosis in Bladder Cancer. *PLOS ONE*, **8**, e59095.
59. Zarei, M., Lal, S., Parker, S.J., Nevler, A., Vaziri-Gohar, A., Dukleska, K., Mambelli-Lisboa, N.C., Moffat, C., Blanco, F.F., Chand, S.N., *et al.* (2017) Posttranscriptional Upregulation of IDH1 by HuR Establishes a Powerful Survival Phenotype in Pancreatic Cancer Cells. *Cancer Research*, **77**, 4460–4471.
60. Lal, S., Cheung, E.C., Zarei, M., Preet, R., Chand, S.N., Mambelli-Lisboa, N.C., Romeo, C., Stout, M.C., Londin, E., Goetz, A., *et al.* (2017) CRISPR Knockout of the HuR Gene Causes a Xenograft Lethal Phenotype. *Molecular Cancer Research*, **15**, 696–707.
61. Blanco, F.F., Jimbo, M., Wulfkuhle, J., Gallagher, I., Deng, J., Enyenihi, L., Meisner-Kober, N., Londin, E., Rigoutsos, I., Sawicki, J.A., *et al.* (2016) The mRNA-binding protein HuR promotes hypoxia-induced chemoresistance through posttranscriptional regulation of the proto-oncogene PIM1 in pancreatic cancer cells. *Oncogene*, **35**, 2529–2541.

62. Mehta,M., Basalingappa,K., Griffith,J.N., Andrade,D., Babu,A., Amreddy,N., Muralidharan,R., Gorospe,M., Herman,T., Ding,W.-Q., *et al.* (2016) HuR silencing elicits oxidative stress and DNA damage and sensitizes human triple-negative breast cancer cells to radiotherapy. *Oncotarget*, **7**, 64820–64835.
63. Sahu,A., Singhal,U. and Chinnaiyan,A.M. (2015) Long Noncoding RNAs in Cancer: From Function to Translation. *Trends in Cancer*, **1**, 93–109.
64. Gutschner,T., Hämmerle,M., Eißmann,M., Hsu,J., Kim,Y., Hung,G., Revenko,A., Arun,G., Stenrup,M., Groß,M., *et al.* (2013) The Noncoding RNA MALAT1 Is a Critical Regulator of the Metastasis Phenotype of Lung Cancer Cells. *Cancer Res*, **73**, 1180–1189.
65. Bhan,A., Soleimani,M. and Mandal,S.S. (2017) Long Noncoding RNA and Cancer: A New Paradigm. *Cancer Research*, **77**, 3965–3981.
66. Arun,G., Diermeier,S., Akerman,M., Chang,K.-C., Wilkinson,J.E., Hearn,S., Kim,Y., MacLeod,A.R., Krainer,A.R., Norton,L., *et al.* (2016) Differentiation of mammary tumors and reduction in metastasis upon Malat1 lncRNA loss. *Genes Dev.*, **30**, 34–51.
67. Gutschner,T., Hämmerle,M. and Diederichs,S. (2013) MALAT1 — a paradigm for long noncoding RNA function in cancer. *J Mol Med*, **91**, 791–801.
68. 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.
69. Baltz,A.G., Munschauer,M., Schwanhäusser,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. *Molecular Cell*, **46**, 674–690.
70. Greenberg,J.R. (1979) Ultraviolet light-induced crosslinking of mRNA to proteins. *Nucleic Acids Research*, **6**, 715–732.
71. Hafner,M., Katsantoni,M., Köster,T., Marks,J., Mukherjee,J., Staiger,D., Ule,J. and Zavolan,M. (2021) CLIP and complementary methods. *Nat Rev Methods Primers*, **1**, 20.
72. Castello,A., Horos,R., Strein,C., Fischer,B., Eichelbaum,K., Steinmetz,L.M., Krijgsveld,J. and Hentze,M.W. (2013) System-wide identification of RNA-binding proteins by interactome capture. *Nature Protocols*, **8**, 491.
73. Conrad,T., Albrecht,A.-S., Costa,V.R. de M., Sauer,S., Meierhofer,D. and Ørom,U.A. (2016) Serial interactome capture of the human cell nucleus. *Nature Communications*, **7**, 11212.
74. 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. *Nature Communications*, **9**, 4408.
75. 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. *Molecular Cell*, **63**, 696–710.
76. 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.

77. 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. *Molecular Cell*, **64**, 416–430.
78. Kolb,H.C., Finn,M.G. and Sharpless,K.B. (2001) Click Chemistry: Diverse Chemical Function from a Few Good Reactions. *Angewandte Chemie International Edition*, **40**, 2004–2021.
79. Bao,X., Guo,X., Yin,M., Tariq,M., Lai,Y., Kanwal,S., Zhou,J., Li,N., Lv,Y., Pulido-Quetglas,C., *et al.* (2018) Capturing the interactome of newly transcribed RNA. *Nature Methods*, 10.1038/nmeth.4595.
80. Huang,R., Han,M., Meng,L. and Chen,X. (2018) Transcriptome-wide discovery of coding and noncoding RNA-binding proteins. *Proceedings of the National Academy of Sciences*, **115**, E3879–E3887.
81. Shchepachev,V., Bresson,S., Spanos,C., Petfalski,E., Fischer,L., Rappsilber,J. and Tollervey,D. (2019) Defining the RNA interactome by total RNA -associated protein purification. *Mol Syst Biol*, **15**.
82. Chomczynski,P. and Sacchi,N. (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Analytical Biochemistry*, **162**, 156–159.
83. 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). *Nature Biotechnology*, 10.1038/s41587-018-0001-2.
84. 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.e19.
85. 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.
86. Scherrer,T., Mittal,N., Janga,S.C. and Gerber,A.P. (2010) A Screen for RNA-Binding Proteins in Yeast Indicates Dual Functions for Many Enzymes. *PLOS ONE*, **5**, e15499.
87. Tsvetanova,N.G., Klass,D.M., Salzman,J. and Brown,P.O. (2010) Proteome-Wide Search Reveals Unexpected RNA-Binding Proteins in *Saccharomyces cerevisiae*. *PLOS ONE*, **5**, e12671.
88. Treiber,T., Treiber,N., Plessmann,U., Harlander,S., Daiß,J.-L., Eichner,N., Lehmann,G., Schall,K., Urlaub,H. and Meister,G. (2017) A Compendium of RNA-Binding Proteins that Regulate MicroRNA Biogenesis. *Molecular Cell*, **66**, 270-284.e13.
89. 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*, 10.1038/s41596-019-0261-4.
90. Mallam,A.L., Sae-Lee,W., Schaub,J.M., Tu,F., Battenhouse,A., Jang,Y.J., Kim,J., Wallingford,J.B., Finkelstein,I.J., Marcotte,E.M., *et al.* (2019) Systematic Discovery of Endogenous Human Ribonucleoprotein Complexes. *Cell Reports*, **29**, 1351-1368.e5.

91. Si,J., Cui,J., Cheng,J. and Wu,R. (2015) Computational Prediction of RNA-Binding Proteins and Binding Sites. *International Journal of Molecular Sciences*, **16**, 26303–26317.
92. Zhang,J., Ma,Z. and Kurgan,L. (2019) Comprehensive review and empirical analysis of hallmarks of DNA-, RNA- and protein-binding residues in protein chains. *Briefings in Bioinformatics*, **20**, 1250–1268.
93. Han,L.Y., Cai,C.Z., Lo,S.L., Chung,M.C.M. and Chen,Y.Z. (2004) Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. *RNA*, **10**, 355–368.
94. Kawashima,S., Pokarowski,P., Pokarowska,M., Kolinski,A., Katayama,T. and Kanehisa,M. (2008) AAindex: amino acid index database, progress report 2008. *Nucleic Acids Research*, **36**, D202–D205.
95. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.
96. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.
97. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Research*, **28**, 235–242.
98. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G., Laydon,A., *et al.* (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, **50**, D439–D444.
99. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Židek,A., Potapenko,A., *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
100. Pepe,G., Appierdo,R., Carrino,C., Ballesio,F., Helmer-Citterich,M. and Gherardini,P. (2022) Artificial intelligence methods enhance the discovery of RNA interactions. *Frontiers in Molecular Biosciences*, **9**.
101. Moreira,I.S., Fernandes,P.A. and Ramos,M.J. (2010) Protein–protein docking dealing with the unknown. *Journal of Computational Chemistry*, **31**, 317–342.
102. 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 Research*, **49**, D605–D612.
103. Brannan,K.W., Jin,W., Huelga,S.C., Banks,C.A.S., 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. *Molecular Cell*, **64**, 282–293.
104. 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. *Nature Communications*, **6**, 10127.

105. Muckenthaler,M.U., Rivella,S., Hentze,M.W. and Galy,B. (2017) A Red Carpet for Iron Metabolism. *Cell*, **168**, 344–361.
106. Hochegger,H., Takeda,S. and Hunt,T. (2008) Cyclin-dependent kinases and cell-cycle transitions: does one fit all? *Nat Rev Mol Cell Biol*, **9**, 910–916.
107. Kanakkanthara,A., Jeganathan,K.B., Limzerwala,J.F., Baker,D.J., Hamada,M., Nam,H.-J., van Deursen,W.H., Hamada,N., Naylor,R.M., Becker,N.A., *et al.* (2016) Cyclin A2 is an RNA binding protein that controls *Mre11* mRNA translation. *Science*, **353**, 1549–1552.
108. 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 Research*, **49**, D425–D436.
109. Lunde,B.M., Moore,C. and Varani,G. (2007) RNA-binding proteins: modular design for efficient function. *Nature Reviews Molecular Cell Biology*, **8**, 479–490.
110. Handa,N., Nureki,O., Kurimoto,K., Kim,I., Sakamoto,H., Shimura,Y., Muto,Y. and Yokoyama,S. (1999) Structural basis for recognition of the tra mRNA precursor by the Sex-lethal protein. *Nature*, **398**, 579–585.
111. Auweter,S.D., Oberstrass,F.C. and Allain,F.H.-T. (2006) Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Research*, **34**, 4943–4959.
112. Valverde,R., Edwards,L. and Regan,L. (2008) Structure and function of KH domains. *The FEBS Journal*, **275**, 2712–2726.
113. Chang,K.-Y. and Ramos,A. (2005) The double-stranded RNA-binding motif, a versatile macromolecular docking platform. *The FEBS Journal*, **272**, 2109–2117.
114. Lu,D., Alexandra Searles,M. and Klug,A. (2003) Crystal structure of a zinc-finger–RNA complex reveals two modes of molecular recognition. *Nature*, **426**, 96–100.
115. Ramos,A., Grünert,S., Adams,J., Micklem,D.R., Proctor,M.R., Freund,S., Bycroft,M., St Johnston,D. and Varani,G. (2000) RNA recognition by a Staufien double-stranded RNA-binding domain. *The EMBO Journal*, **19**, 997–1009.
116. Ha,M. and Kim,V.N. (2014) Regulation of microRNA biogenesis. *Nat Rev Mol Cell Biol*, **15**, 509–524.
117. MacRae,I.J., Zhou,K., Li,F., Repic,A., Brooks,A.N., Cande,W.Z., Adams,P.D. and Doudna,J.A. (2006) Structural Basis for Double-Stranded RNA Processing by Dicer. *Science*, **311**, 195–198.
118. Jinek,M. and Doudna,J.A. (2009) A three-dimensional view of the molecular machinery of RNA interference. *Nature*, **457**, 405–412.
119. Liu,Z., Luyten,I., Bottomley,M.J., Messias,A.C., Houngninou-Molango,S., Sprangers,R., Zanier,K., Krämer,A. and Sattler,M. (2001) Structural Basis for Recognition of the Intron Branch Site RNA by Splicing Factor 1. *Science*, **294**, 1098–1102.
120. Neelamraju,Y., Hashemikhabir,S. and Janga,S.C. (2015) The human RBPome: From genes and proteins to human disease. *Journal of Proteomics*, **127**, 61–70.
121. Schopf,F.H., Biebl,M.M. and Buchner,J. (2017) The HSP90 chaperone machinery. *Nat Rev Mol Cell Biol*, **18**, 345–360.

122. Prodromou,C., Roe,S.M., Piper,P.W. and Pearl,L.H. (1997) A molecular clamp in the crystal structure of the N-terminal domain of the yeast Hsp90 chaperone. *Nat Struct Mol Biol*, **4**, 477–482.
123. Panhale,A., Richter,F.M., Ramírez,F., Shvedunova,M., Manke,T., Mittler,G. and Akhtar,A. (2019) CAPRI enables comparison of evolutionarily conserved RNA interacting regions. *Nat Commun*, **10**, 2682.
124. Iwasaki,S., Kobayashi,M., Yoda,M., Sakaguchi,Y., Katsuma,S., Suzuki,T. and Tomari,Y. (2010) Hsc70/Hsp90 Chaperone Machinery Mediates ATP-Dependent RISC Loading of Small RNA Duplexes. *Molecular Cell*, **39**, 292–299.
125. Kishor,A., Tandukar,B., Ly,Y.V., Toth,E.A., Suarez,Y., Brewer,G. and Wilson,G.M. (2013) Hsp70 Is a Novel Posttranscriptional Regulator of Gene Expression That Binds and Stabilizes Selected mRNAs Containing AU-Rich Elements. *Molecular and Cellular Biology*, **33**, 71–84.
126. Kishor,A., White,E.J.F., Matsangos,A.E., Yan,Z., Tandukar,B. and Wilson,G.M. (2017) Hsp70's RNA-binding and mRNA-stabilizing activities are independent of its protein chaperone functions. *Journal of Biological Chemistry*, **292**, 14122–14133.
127. Järvelin,A.I., Noerenberg,M., Davis,I. and Castello,A. (2016) The new (dis)order in RNA regulation. *Cell Commun Signal*, **14**, 9.
128. Hyman,A.A., Weber,C.A. and Jülicher,F. (2014) Liquid-Liquid Phase Separation in Biology. *Annu. Rev. Cell Dev. Biol.*, **30**, 39–58.
129. Basu,S. and Bahadur,R.P. (2016) A structural perspective of RNA recognition by intrinsically disordered proteins. *Cell. Mol. Life Sci.*, **73**, 4075–4084.
130. Stefl,R., Xu,M., Skrisovska,L., Emeson,R.B. and Allain,F.H.-T. (2006) Structure and Specific RNA Binding of ADAR2 Double-Stranded RNA Binding Motifs. *Structures*, 10.1016/j.str.2005.11.013.
131. Smith,C.A., Calabro,V. and Frankel,A.D. (2000) An RNA-Binding Chameleon. *Molecular Cell*, **6**, 1067–1076.
132. Strein,C., Alleaume,A.-M., Rothbauer,U., Hentze,M.W. and Castello,A. (2014) A versatile assay for RNA-binding proteins in living cells. *RNA*, **20**, 721–731.
133. Kiledjian,M. and Dreyfuss,G. (1992) Primary structure and binding activity of the hnRNP U protein: binding RNA through RGG box. *The EMBO Journal*, **11**, 2655–2664.
134. Thandapani,P., O'Connor,T.R., Bailey,T.L. and Richard,S. (2013) Defining the RGG/RG Motif. *Molecular Cell*, **50**, 613–623.
135. Hasegawa,Y., Brockdorff,N., Kawano,S., Tsutui,K., Tsutui,K. and Nakagawa,S. (2010) The Matrix Protein hnRNP U Is Required for Chromosomal Localization of Xist RNA. *Developmental Cell*, **19**, 469–476.
136. 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.

137. Xiang,S., Gapsys,V., Kim,H.-Y., Bessonov,S., Hsiao,H.-H., Möhlmann,S., Klaukien,V., Ficner,R., Becker,S., Urlaub,H., *et al.* (2013) Phosphorylation Drives a Dynamic Switch in Serine/Arginine-Rich Proteins. *Structure*, **21**, 2162–2174.
138. Loughlin,F.E., Lukavsky,P.J., Kazeeva,T., Reber,S., Hock,E.-M., Colombo,M., Von Schroetter,C., Pauli,P., Cléry,A., Mühlemann,O., *et al.* (2019) The Solution Structure of FUS Bound to RNA Reveals a Bipartite Mode of RNA Recognition with Both Sequence and Shape Specificity. *Molecular Cell*, **73**, 490-504.e6.
139. 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.
140. Zhao,B., Katuwawala,A., Oldfield,C.J., Hu,G., Wu,Z., Uversky,V.N. and Kurgan,L. (2021) Intrinsic Disorder in Human RNA-Binding Proteins. *Journal of Molecular Biology*, **433**, 167229.
141. Linding,R., Jensen,L.J., Diella,F., Bork,P., Gibson,T.J. and Russell,R.B. (2003) Protein Disorder Prediction: Implications for Structural Proteomics. *Structure*, **11**, 1453–1459.
142. 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.
143. 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.
144. The UniProt Consortium, Bateman,A., Martin,M.-J., Orchard,S., Magrane,M., Agivetova,R., Ahmad,S., Alpi,E., Bowler-Barnett,E.H., Britto,R., *et al.* (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, **49**, D480–D489.
145. Blum,M., Chang,H.-Y., Chuguransky,S., Grego,T., Kandasamy,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 Research*, **49**, D344–D354.
146. Monastyrskyy,B., Kryshtafovych,A., Moulton,J., Tramontano,A. and Fidelis,K. (2014) Assessment of protein disorder region predictions in CASP10. *Proteins: Structure, Function, and Bioinformatics*, **82**, 127–137.
147. CAID Predictors, DisProt Curators, Necci,M., Piovesan,D. and Tosatto,S.C.E. (2021) Critical assessment of protein intrinsic disorder prediction. *Nat Methods*, **18**, 472–481.
148. Gerstberger,S., Hafner,M. and Tuschl,T. (2014) A census of human RNA-binding proteins. *Nature Reviews Genetics*, **15**, 829.
149. Burge,S., Kelly,E., Lonsdale,D., Mutowo-Muellenet,P., McAnulla,C., Mitchell,A., Sangrador-Vegas,A., Yong,S.-Y., Mulder,N. and Hunter,S. (2012) Manual GO annotation of predictive protein signatures: the InterPro approach to GO curation. *Database*, **2012**, bar068–bar068.
150. Bekker-Jensen,D.B., Kelstrup,C.D., Bath,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.e4.

151. Cook,K.B., Kazan,H., Zuberi,K., Morris,Q. and Hughes,T.R. (2011) RBPDB: a database of RNA-binding specificities. *Nucleic Acids Research*, **39**, D301–D308.
152. Burley,S.K., Bhikadiya,C., Bi,C., Bittrich,S., Chao,H., Chen,L., Craig,P.A., Crichlow,G.V., Dalenberg,K., Duarte,J.M., *et al.* (2023) RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. *Nucleic Acids Research*, **51**, D488–D508.
153. Das,R.K. and Pappu,R.V. (2013) Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 13392–13397.
154. Ford,L.K. and Fioriti,L. (2020) Coiled-Coil Motifs of RNA-Binding Proteins: Dynamicity in RNA Regulation. *Front. Cell Dev. Biol.*, **8**, 607947.
155. 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.
156. Malarkey,C.S. and Churchill,M.E.A. (2012) The high mobility group box: the ultimate utility player of a cell. *Trends in Biochemical Sciences*, **37**, 553–562.
157. Hamilton,D.J., Hein,A.E., Wuttke,D.S. and Batey,R.T. (2023) The DNA-binding high mobility group box protein family functionally binds RNA. *WIREs RNA*, 10.1002/wrna.1778.
158. Xue,B., Dunker,A.K. and Uversky,V.N. (2012) Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *Journal of Biomolecular Structure and Dynamics*, **30**, 137–149.
159. Schad,E., Tompa,P. and Hegyi,H. (2011) The relationship between proteome size, structural disorder and organism complexity. *Genome Biology*, **12**, R120.
160. Jain,B.P. and Pandey,S. (2018) WD40 Repeat Proteins: Signalling Scaffold with Diverse Functions. *Protein J*, **37**, 391–406.
161. Bi,X., Xu,Y., Li,T., Li,X., Li,W., Shao,W., Wang,K., Zhan,G., Wu,Z., Liu,W., *et al.* (2019) RNA Targets Ribogenesis Factor WDR43 to Chromatin for Transcription and Pluripotency Control. *Molecular Cell*, **75**, 102-116.e9.
162. Johnson,B., VanBlargan,L.A., Xu,W., White,J.P., Shan,C., Shi,P.-Y., Zhang,R., Adhikari,J., Gross,M.L., Leung,D.W., *et al.* (2018) Human IFIT3 Modulates IFIT1 RNA Binding Specificity and Protein Stability. *Immunity*, **48**, 487-499.e5.
163. Katibah,G.E., Qin,Y., Sidote,D.J., Yao,J., Lambowitz,A.M. and Collins,K. (2014) Broad and adaptable RNA structure recognition by the human interferon-induced tetratricopeptide repeat protein IFIT5. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 12025–12030.
164. Rajiv,C. and Davis,T. (2018) Structural and Functional Insights into Human Nuclear Cyclophilins. *Biomolecules*, **8**, 161.
165. Chang,A.-Y., Castel,S.E., Ernst,E., Kim,H.S. and Martienssen,R.A. (2017) The Conserved RNA Binding Cyclophilin, Rct1, Regulates Small RNA Biogenesis and Splicing Independent of Heterochromatin Assembly. *Cell Reports*, **19**, 2477–2489.

166. Dias,S.M.G., Cerione,R.A. and Wilson,K.F. (2010) Unloading RNAs in the cytoplasm: An “importin” task. *Nucleus*, **1**, 139–143.
167. Huang,Y., Jiang,Z., Gao,X., Luo,P. and Jiang,X. (2021) ARMC Subfamily: Structures, Functions, Evolutions, Interactions, and Diseases. *Front. Mol. Biosci.*, **8**, 791597.
168. Ray,D., Laverty,K.U., Jolma,A., Nie,K., Samson,R., Pour,S.E., Tam,C.L., von Krosigk,N., Nabeel-Shah,S., Albu,M., *et al.* (2023) RNA-binding proteins that lack canonical RNA-binding domains are rarely sequence-specific. *Sci Rep*, **13**, 5238.
169. Lee,F.C.Y. and Ule,J. (2018) Advances in CLIP Technologies for Studies of Protein-RNA Interactions. *Molecular Cell*, **69**, 354–369.
170. Harris,M.E. and Christian,E.L. (2009) Chapter 7 - RNA Crosslinking Methods. In *Methods in Enzymology*, Biophysical, Chemical, and Functional Probes of RNA Structure, Interactions and Folding: Part A. Academic Press, Vol. 468, pp. 127–146.
171. Mellacheruvu,D., Wright,Z., Couzens,A.L., Lambert,J.-P., St-Denis,N.A., Li,T., Miteva,Y.V., Hauri,S., Sardu,M.E., Low,T.Y., *et al.* (2013) The CRAPome: a contaminant repository for affinity purification–mass spectrometry data. *Nat Methods*, **10**, 730–736.
172. Riemondy,K.A., Sheridan,R.M., Gillen,A., Yu,Y., Bennett,C.G. and Hesselberth,J.R. (2017) valr: Reproducible genome interval analysis in R. *F1000Res*, **6**, 1025.
173. Liu,Y., Wang,X. and Liu,B. (2019) A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. *Briefings in Bioinformatics*, **20**, 330–346.
174. Quaglia,F., Mészáros,B., Salladini,E., Hatos,A., Pancsa,R., Chemes,L.B., Pajkos,M., Lazar,T., Peña-Díaz,S., Santos,J., *et al.* (2022) DisProt in 2022: improved quality and accessibility of protein intrinsic disorder annotation. *Nucleic Acids Research*, **50**, D480–D487.
175. Fukuchi,S., Sakamoto,S., Nobe,Y., Murakami,S.D., Amemiya,T., Hosoda,K., Koike,R., Hiroaki,H. and Ota,M. (2012) IDEAL: Intrinsically Disordered proteins with Extensive Annotations and Literature. *Nucleic Acids Research*, **40**, D507–D511.
176. Oates,M.E., Romero,P., Ishida,T., Ghalwash,M., Mizianty,M.J., Xue,B., Dosztányi,Z., Uversky,V.N., Obradovic,Z., Kurgan,L., *et al.* (2013) D2P2: database of disordered protein predictions. *Nucleic Acids Research*, **41**, D508–D516.
177. Rezwani,M., Pourfathollah,A.A. and Noorbakhsh,F. (2022) rbioapi: user-friendly R interface to biologic web services’ API. *Bioinformatics*, **38**, 2952–2953.
178. Hadley Wickham (2016) ggplot2: Elegant Graphics for Data Analysis Springer-Verlag New York.
179. 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.
180. Liao,Y., Castello,A., Fischer,B., Leicht,S., Föehr,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 Reports*, **16**, 1456–1469.

181. 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. *The Plant Cell*, **28**, 2435–2452.
182. Szklarczyk,D., Gable,A.L., Lyon,D., Junge,A., Wyder,S., Huerta-Cepas,J., Simonovic,M., Doncheva,N.T., Morris,J.H., Bork,P., *et al.* (2019) STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, **47**, D607–D613.
183. R Core Team (2021) R: A Language and Environment for Statistical Computing.