

Mass Spectrometry-based Quantitative Proteomics to Investigate RNA-Protein Interactions

Dissertation

zur Erlangung des Grades

„Doktor der Naturwissenschaften“

am Fachbereich Biologie

der Johannes Gutenberg-Universität in Mainz

Albert Fradera-Sola

geb. am 27.10.1990 in Barcelona, Spanien

Mainz, Januar 2024



Dekan: Prof. Dr. Eckhard Thines

1. Berichterstatter: Prof. Dr. Miguel Andrade
2. Berichterstatter: Dr. Falk Butter
3. Betreuer: Prof. Dr. Miguel Andrade
4. Betreuer: Prof. Dr. Susanne Gerber
5. Vorsitz: Prof. Dr. Eva Wolf
6. Protokollant: Dr. Sabrina Dietz

Tag der mündlichen Prüfung: 19.01.2024

Abstract

Mass spectrometry-based proteomics is a versatile tool, offering a global and unbiased approach to analyse proteins and their interacting partners. Within the realm of molecular biology, RNA-protein interactions stand as fundamental and intricate components that oversee vital processes in the cell. These interactions, often mediated by specific RNA-binding proteins (RBPs), orchestrate a wide array of cellular functions. From the regulation of gene expression to the maintenance of genomic stability, the post-transcriptional processing of RNA molecules, and even the spatial organisation of the cell nucleus, RNA-protein interactions play a central role in shaping the intricate web of cellular activities. In this thesis, the interaction between RNA and proteins is investigated using state-of-the-art MS.

Article I delved into the characterization of Telomeric repeat-containing RNA (TERRA) molecules in *M. musculus*, examining their genomic origins and comparing their interactomes to their *H. sapiens* counterparts. RNA-FISH analysis revealed disparities in behaviour, with *M. musculus* TERRA foci primarily located outside of telomeres, in contrast to *H. sapiens* TERRA foci, which recurrently resided at telomeres. As a result, a distinct genomic origin for *M. musculus* TERRA molecules outside telomeres was hypothesised. Through a comprehensive genomic analysis, four major chromosomal regions, including known Telo 18q, PAR-Xq/Yq, and ChrX Tsix locus regions, and a novel Chr2 region, were identified as potential sources of TERRA molecules. Conservation of TERRA-associated functions was evaluated with an affinity purification-mass spectrometry (AP-MS) approach. A comparison of the enriched proteins with publicly available *H. sapiens* TERRA-interacting protein datasets revealed that, despite having a distinct genomic origin, functions are conserved between *M. musculus* and *H. sapiens*.

Article II centred on the functional assignment of RBPs in *S. cerevisiae*. An AP-MS screen was designed to elucidate the interaction partners of 40 selected RBPs, which were chosen based on their involvement in various stages of mRNA processing. Functional analysis of the collected data highlighted the overrepresentation of canonical RNA-binding domains (RBDs) and RNA binding-related GO molecular function terms among the RBPs' interaction partners. KEGG pathway analysis demonstrated the enrichment of RNA pathways, consistent with the RBP selection criteria, as well as involvement in metabolic and synthesis pathways. Finally, network-based function assignment of RBPs was facilitated by concurrent binding patterns within the network.

Zusammenfassung

Auf Massenspektrometrie Proteomik ist ein vielseitiges Werkzeug, das einen globalen und unvoreingenommenen Ansatz zur Analyse von Proteinen und deren Interaktionspartnern bietet. Im Bereich der Molekularbiologie stellen RNA-Protein-Interaktionen grundlegende und komplexe Komponenten dar, die für wichtige Prozesse in der Zelle notwendig sind. Diese Interaktionen, die oft von spezifischen RNA-bindenden Proteinen (RBPs) vermittelt werden, orchestrieren eine Vielzahl zellulärer Funktionen. Von der Regulation der Genexpression über die Aufrechterhaltung der genomischen Stabilität, die post-transkriptionelle Verarbeitung von RNA-Molekülen bis hin zur räumlichen Organisation des Zellkerns spielen RNA-Protein-Interaktionen eine zentrale Rolle bei der Gestaltung des komplexen Geflechts zellulärer Aktivitäten. In dieser Dissertation wird die Interaktion zwischen RNA und Proteinen mithilfe modernster Massenspektrometrie untersucht.

Artikel I befasste sich mit der Charakterisierung von Telomeric Repeat-Containing RNA (TERRA)-Molekülen in *M. musculus*, untersuchte ihre genomischen Ursprünge und verglich ihre Interaktome mit ihren *H. sapiens*-Gegenstücken. Die RNA-FISH-Analyse zeigte Unterschiede im Verhalten auf, wobei sich die TERRA-Foki von *M. musculus* hauptsächlich außerhalb der Telomere befanden, im Gegensatz zu den TERRA-Foki von *H. sapiens*, die wiederholt an den Telomeren zu finden waren. Basierend darauf wurde eine unterschiedliche genomische Herkunft der TERRA-Moleküle von *M. musculus* außerhalb der Telomere vermutet. Durch eine umfassende genomische Analyse wurden vier Hauptchromosomenregionen, darunter bekannte Telo 18q, PAR-Xq/Yq und ChrX Tsix-Locus-Regionen sowie eine neue Chr2-Region, identifiziert, die als potenzielle Quellen für TERRA-Moleküle in Frage kommen. Die Konservierung der mit TERRA assoziierten Funktionen wurde mithilfe eines Affinitätsreinigung-Massenspektrometrie (AP-MS)-Ansatzes bewertet. Ein Vergleich der angereicherten Proteine mit öffentlich zugänglichen Datensätzen von *H. sapiens* TERRA-interagierenden Proteinen zeigte, dass trotz einer unterschiedlichen genomischen Herkunft die Funktionen zwischen *M. musculus* und *H. sapiens* erhalten bleiben.

Artikel II handelt von der funktionalen Zuordnung von RBPs in *S. cerevisiae*. Es wurde ein AP-MS-Screen entwickelt, um die Interaktionspartner von 40 ausgewählten RBPs zu ermitteln, die aufgrund ihrer Beteiligung an verschiedenen Phasen der mRNA-Prozessierung ausgewählt wurden. Die funktionale Analyse der gesammelten Daten ergab eine Überrepräsentation von kanonischen RNA-bindenden Domänen (RBDs) und von GO-Begriffen zu molekularen Funktionen für RNA-Bindungen unter den Interaktionspartnern der RBPs. Die

KEGG-Weg -Analyse zeigte, entsprechend dem Selektionskriterium für RBPs, eine Anreicherung von RNA-Wegen , sowie deren Beteiligung an Stoffwechsel- und Synthesewegen. Abschließend wurde die netzwerkbasierete funktionale Zuordnung von RBPs durch gleichzeitige Bindungsmuster im Netzwerk ermöglicht .

List of publications

Thesis publications

Article I :

Viceconte, N., Lorient, A., Lona Abreu, P., Scheibe, M., Fradera Sola, A., Butter, F., De Smet, C., Azzalin, C. M., Arnoult, N., & Decottignies, A. (2021). PAR-TERRA is the main contributor to telomeric repeat-containing RNA transcripts in normal and cancer mouse cells. *RNA*, 27(1), 106–121. <https://doi.org/10.1261/rna.076281.120>

Article II :

Fradera-Sola, A., Nischwitz, E., Bayer, M. E., Luck, K., & Butter, F. (2023). RNA-dependent interactome allows network-based assignment of RNA-binding protein function. *Nucleic Acids Research*, gkad245. <https://doi.org/10.1093/nar/gkad245>

Additional publications

Amodeo, S., Kalichava, A., Fradera-Sola, A., Bertiaux-Lequoy, E., Guichard, P., Butter, F., & Ochsenreiter, T. (2021). Characterization of the novel mitochondrial genome segregation factor TAP110 in *Trypanosoma brucei*. *Journal of Cell Science*, 134(5), jcs254300. <https://doi.org/10.1242/jcs.254300>

Dietz, S., Almeida, M. V., Nischwitz, E., Schreier, J., Viceconte, N., Fradera-Sola, A., Renz, C., Ceron-Noriega, A., Ulrich, H. D., Kappei, D., Ketting, R. F., & Butter, F. (2021). The double-stranded DNA-binding proteins TEBP-1 and TEBP-2 form a telomeric complex with POT-1. *Nature Communications*, 12(1), 2668. <https://doi.org/10.1038/s41467-021-22861-2>

- Vellmer, T., Hartleb, L., Fradera Sola, A., Kramer, S., Meyer-Natus, E., Butter, F., & Janzen, C. J. (2022). A novel SNF2 ATPase complex in *Trypanosoma brucei* with a role in H2A.Z-mediated chromatin remodelling. *PLOS Pathogens*, 18(6), e1010514. <https://doi.org/10.1371/journal.ppat.1010514>
- Amodeo, S., Bregy, I., Hoffmann, A., Fradera-Sola, A., Kern, M., Baudouin, H., Zuber, B., Butter, F., & Ochsenreiter, T. (2023). Characterization of two novel proteins involved in mitochondrial DNA anchoring in *Trypanosoma brucei*. *PLOS Pathogens*, 19(7), e1011486. <https://doi.org/10.1371/journal.ppat.1011486>
- Nischwitz, E., Schoonenberg, V. A. C., Fradera-Sola, A., Dejung, M., Vydzhak, O., Levin, M., Luke, B., Butter, F., & Scheibe, M. (2023). DNA damage repair proteins across the Tree of Life. *iScience*, 26(6), 106778. <https://doi.org/10.1016/j.isci.2023.106778>

Abbreviations

ACN	Acetonitrile
AP	Affinity purification
AQUA	Absolute quantification
Aurkb	Aurora kinase B
CID	Collision-induced dissociation
CV	Compensation voltage
DDA	Data dependent acquisition
DIA	Data independent acquisition
DTT	Dithiothreitol
emPAI	Exponentially modified protein abundance index
ESC	Embryonic stem cells
ESI	Electrospray ionisation
FAIMS	Field asymmetric waveform ion mobility spectrometry
FISH	Fluorescence in situ hybridisation
FP	False positive
GO	Gene ontology
HCD	Higher-energy C-trap dissociation
HPLC	High performance liquid chromatography
IAA	2-iodoacetamidaurkbde
IMS	Ion mobility spectrometry
iDRiP	Identification of direct RNA interacting proteins
IP	Immunoprecipitation
iTRAQ	Isobaric tag for relative and absolute quantification
KO	Knockout
lncRNA	Long non-coding RNA
MALDI	Matrix-assisted laser desorption/ionisation
MAR	Missing at random
MCAR	Missing completely at random
miRNA	MicroRNAs
MinProb	Probabilistic minimum imputation
MNAR	Missing not at random
mRNA	Messenger RNA

MS	Mass spectrometry
m/z	Mass-to-charge ratio
NGS	Next generation sequencing
PAI	Protein abundance index
PAR	Pseudoautosomal region
PCA	Principal component analysis
Pfam	Protein families
PPI	Protein-protein interaction
PSMs	Peptide-spectra matches
QTOF	Quadrupole-time of flight
R	Residue, amino acid side chain
RBD	RNA binding domain
RBP	RNA binding protein
RIC	RNA interactome capture
RDI	RNA dependant interactions
RINE	RBP interactome network explorer
RNP	Ribonucleoprotein
rRNA	Ribosomal RNA
RT-qPCR	Reverse-transcription real-time PCR
SDS-PAGE	Sodium dodecyl sulphate–polyacrylamide gel electrophoresis
SILAC	Stable isotope labelling by amino acids in cell culture
snRNA	Small nuclear RNA
snRNP	Small nuclear RNP
SOM	Self-organising map
StageTips	Stop-and-go-extraction tips
TERRA	Telomeric repeat-containing RNA
TIMS	Trapped ion mobility spectrometry
TMT	Tandem mass tag
TOF	Time-of-flight
TREX	Transcription export complex
tRNA	Transfer RNA
UV	Ultraviolet
XIC	Extracted ion current
Xic	X-inactivation centre

Table of contents

Abstract	i
Zusammenfassung	ii
List of publications	iv
Thesis publications.....	iv
Additional publications.....	iv
Abbreviations	vi
1. Introduction	1
1.1 Mass spectrometry-based proteomics.....	1
1.1.1 Top-down proteomics.....	2
1.1.2 Bottom-up proteomics.....	2
1.2 Sample processing.....	4
1.2.1 Sample fractionation and protein purification.....	4
1.2.2 Protein digestion.....	4
1.2.3 Peptide separation.....	5
1.2.4 Peptide ionisation.....	6
1.2.4.1 MALDI.....	6
1.2.4.2 ESI.....	8
1.2.5 Ion mobility spectrometry.....	8
1.2.5.1 TIMS.....	8
1.2.5.2 FAIMS.....	9
1.3 Mass spectrometer.....	10
1.3.1 Quadrupole - Time of flight.....	11
1.3.2 Quadrupole - Orbitrap.....	13
1.4 Data acquisition strategies.....	13
1.5 Mass spectrometry spectra.....	15
1.6 Data analysis.....	16
1.6.1 Peptide identification.....	17
1.6.2 Quantification strategies.....	18
1.6.2.1 Label-based quantification strategies.....	18
1.6.2.2 Label-free quantification strategies.....	20
1.6.3 Missing value imputation.....	21
1.6.4 Exploratory data analysis.....	22
1.6.5 Statistical difference assessment.....	24
1.6.6 Functional analysis.....	25
1.7 Applications to study RNA-protein interactions.....	26
1.7.1 Telomeric repeat-containing RNA origin and interacting partners.....	28
1.7.2 RNA binding protein network-based function assignment.....	28
2. Aims of the thesis	30

3. Articles.....	31
3.1. PAR-TERRA is the main contributor to telomeric repeat-containing RNA transcripts in normal and cancer mouse cells.....	31
3.1.1 Summary.....	31
3.1.2 Zusammenfassung.....	32
3.1.3 Statement of contribution.....	33
3.2 RNA-dependent interactome allows network-based assignment of RNA-binding protein function.....	58
3.2.1 Summary.....	58
3.2.2 Zusammenfassung.....	59
3.2.3 Statement of contribution.....	60
4. Conclusions and future perspectives.....	84
4.1 Telomeric repeat-containing RNA origin and interacting partners.....	84
4.2 RNA binding protein network-based function assignment.....	87
4.3 MS-based proteomics to study RNA-protein interactions.....	91
5. References.....	93
Acknowledgements.....	104
Curriculum vitae.....	106

1. Introduction

Proteins are one of the most important functional effectors of the cell. They constitute about 50% of a cell's dry mass and can reach total concentrations of 2–4 million proteins per cubic micrometre (Milo, 2013). For instance, a *Schizosaccharomyces pombe* cell contains around 60 million protein molecules, with their expression levels ranging from a few to a million copies per gene (Marguerat et al., 2012). Collectively, these proteins constitute the proteome, which, essentially, drives and regulates all biological processes. Across the tree of life, proteomes vary in size, with model organisms such as *Saccharomyces cerevisiae*, *Danio rerio*, and *Homo sapiens* having approximately 5,000, 9,000, and 12,000 protein-coding genes, respectively (Müller et al., 2020).

The proteome is a dynamic entity that adjusts in response to both internal and external perturbations, reflecting the functional state of a biological system, whether it is a cell, tissue, or organism (Aebersold & Mann, 2016). Furthermore, the abundance of proteins can be just as important as their absence or presence when comparing different biological states within the same model organism (de Godoy et al., 2008; Lundberg et al., 2010). Therefore, accurately identifying and quantifying proteomes and their dynamics is crucial for understanding the functional state of a biological system across a wide range of fields.

1.1 Mass spectrometry-based proteomics

Mass spectrometry (MS) is an analytical technique used to detect the presence and abundance of analytes based on fundamental properties of molecules, such as mass and net charge. In MS-based proteomics, proteins are first extracted from a sample, separated, and then ionised into gas-phase ions. These ions are subsequently introduced into a mass spectrometer, where they are separated according to their mass-to-charge ratio (m/z) and detected for their abundance. The resulting mass spectrum provides information on the mass and relative abundance of the protein ions in the sample. Compared to classical biochemical methodologies, MS allows for large-scale systematic protein measurement in an unbiased fashion (Aebersold & Mann, 2003).

To identify the proteins in the sample, the mass spectra are compared to a database of known protein sequences. This is accomplished using software programs that match the experimental mass spectra to theoretical mass spectra generated from protein sequence

databases (Steen & Mann, 2004). Advances in MS proteomics databases have been fueled by the rise of next-generation sequencing (NGS) and the expansion of available genome data sets; genomic and transcriptomic information is used to generate and expand the protein sequence databases, maximising protein identifications (Ceron-Noriega et al., 2023; Nesvizhskii, 2014).

MS-based proteomics can also be used for protein quantification. This is achieved by measuring the abundance of the identified peptides and inferring the abundance of the corresponding proteins. Typically, this is done by comparing the abundance of peptides between two or more samples, or by using labelling techniques (Steen & Mann, 2004).

1.1.1 Top-down proteomics

In top-down proteomics, intact proteins, without prior digestion into smaller peptides, are ionised and then transferred into the mass spectrometer. Subsequently, they are fragmented and analysed to obtain information about their primary structure, post-translational modifications, and other properties. Top-down proteomics offers unique advantages over the more commonly used bottom-up proteomics approach. By interrogating intact proteins, in principle, a complete sequence coverage is achieved. This enables the ability to identify and quantify protein isoforms and modifications that may be missed by bottom-up approaches (Tran et al., 2011). However, it has major limitations; intact, large proteins are not efficiently dissociated, which generally restricts this approach to smaller proteins (<10 kDa) (Shaw et al., 2013). Moreover, top-down proteomics remains a challenging technique, both experimentally and computationally, due to the difficulty of analysing full proteins instead of peptides, limiting the proteome coverage in this manner (Aebersold & Mann, 2016).

1.1.2 Bottom-up proteomics

In bottom-up proteomics, peptides generated by enzymatic digestion of proteins are ionised and then transferred into the mass spectrometer, where a first mass spectrum is recorded. The ions are then fragmented, and a second spectrum is recorded. These spectra are later used to identify and quantify specific peptides (Figure 1). This setup is called tandem MS or MS/MS (MS^2), and its main advantage is its high sensitivity and throughput. It enables the detection and identification of low-abundance proteins and peptides in complex mixtures for multiple samples. There are two main approaches in bottom-up proteomics: shotgun proteomics, aimed at achieving an unbiased and complete proteome coverage, either with a data dependent

acquisition (DDA) or a data independent acquisition (DIA) strategy, and targeted proteomics, aimed to monitor a subset of known peptides of interest (Aebersold & Mann, 2016).

In DDA shotgun proteomics, a popular configuration for the mass spectrometer is a quadrupole-orbitrap analyser. This kind of instrument alternates between acquiring mass spectra at the MS¹ level (precursor ion scan) and at the MS² level (product ion scan). This way, MS¹ spectra are first recorded for all ionised peptides co-eluting at a specific time point in the gradient. Then, top N cycles (N indicating the number of MS² spectra that follow an MS¹ spectrum) follow to acquire the mass spectra at the MS² level for the fragmented ions (Aebersold & Mann, 2016). On the other hand, in DIA-based proteomics, entire ranges of precursor ions are fragmented simultaneously. This results in a multiplexed MS² spectrum whose complexity is usually disentangled by targeted signal extraction (based on previously acquired single-peptide fragmentation spectra library). With this approach, the entire range of possible precursor-ion masses is analysed (Aebersold & Mann, 2016; Chapman et al., 2014).

In targeted proteomics, the mass spectrometer configuration usually consists of a triple quadrupole analyser. The proteins of interest are predetermined, as are their subset of peptides, which instruct the MS measurement. The first quadrupole is set to acquire spectra within the expected precursor ion m/z range, the second quadrupole serves as a collision chamber to generate fragmented ions, and the third quadrupole is set to the m/z ratio of a particularly abundant fragment ion specific to the targeted peptide. The process can be multiplexed to several fragments per peptide with a technique called multiple reaction monitoring (MRM) (Aebersold & Mann, 2016).

Although each experimental setup requires specific fine-tuning, there are some general steps and guidelines applicable to most proteomic sample preparation experiments. These are discussed in the following section, with a particular focus on bottom-up proteomics.

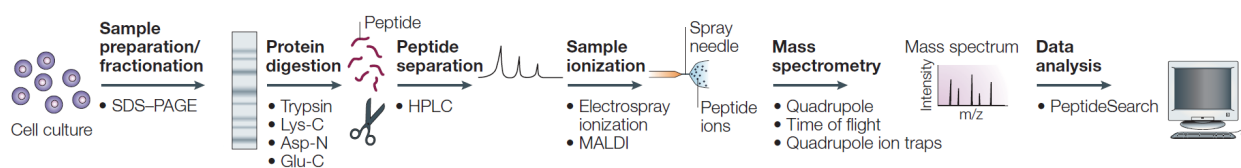


Figure 1. Bottom-up proteomics experimental setup. Proteins are extracted from a biological source and digested into peptides. The peptide mixture is separated by hydrophobicity with an HPLC system. Eluted peptides are ionised and can be analysed with various mass spectrometers, which generate a mass spectrum displaying the m/z ratio (x-axis) and intensity (y-axis) per ion. Finally, the generated mass spectra are searched against protein databases to enable peptide sequence identification and quantification. HPLC, high-performance liquid chromatography. Modified from (Steen & Mann, 2004).

1.2 Sample processing

1.2.1 Sample fractionation and protein purification

The proteome is initially extracted from its biological source, such as a cell culture, relying on its physicochemical characteristics. The extraction process begins with a homogenisation step, which includes various methods such as mechanical, ultrasonic, pressure, freeze-thaw and osmotic/detergent lysis techniques. As a consequence of the homogenisation procedure, the sample's physical properties change, while its chemical properties remain unchanged. With the disruption of cell walls and membranes, proteins are isolated but are generally found in their native state: insoluble and associated with other proteins or membranes. Therefore, protein aggregation must be disrupted to solubilize the proteins. To achieve this, buffers containing detergents, typically within a concentration range of 1%-4%, are used to prevent hydrophobic interactions (Ahmed, 2009; Mostovenko et al., 2013).

In addition to detergents, the buffer may include various reagents, either individually or in combination, to aid in solubilization and protein denaturation. These reagents include: (1) chaotropes, which disrupt hydrogen bonds and hydrophilic interactions; (2) reducing reagents, which break disulfide bonds between cysteines, promoting protein unfolding; and (3) protease inhibitors, which protect the samples from proteolysis. Finally, a last step involves removing contaminants and purifying the proteins, often achieved through the SDS-PAGE methodology (Ahmed, 2009; Mostovenko et al., 2013).

1.2.2 Protein digestion

In bottom-up proteomics, proteins need to be digested into peptides prior to measurement and quantification (Figure 2). When coupled with protein fractionation and purification using in-gel digestion, the resulting SDS-PAGE is chopped into pieces, and the digestion is performed within the gel. Otherwise, it is performed in-solution. Regardless of the strategy, proteins' disulfide bonds are disrupted with reducing reagents such as dithiothreitol (DTT). Then, the newly created sulfhydryl group (SH) is protected from further oxidation by using alkylating reagents such as 2-iodoacetamide (IAA) or chloroacetamide. At this point, proteins are digested with site-specific proteases, enzymes that catalyse proteolysis, breaking down the proteins into smaller polypeptides at specific amino acid residues (Shevchenko et al., 2006).

Sequence specific proteases are widely used because their digestion product is a less complex mixture. This helps with peptide identification later on, as the peptide signal is not divided into overlapping species (Steen & Mann, 2004). Trypsin, for instance, is a widely used protease that cleaves proteins on the carboxy-terminal side of arginine and lysine residues, unless followed by a proline residue. Not only is it sequence specific, but it is also a very active and stable enzyme, maximising its efficiency, hence its popularity. Additionally, peptides digested with trypsin end up with a positive charge at the peptide C-terminus, which is advantageous for MS analysis (Olsen et al., 2004). Another popular enzyme is the endoproteinase Lys-C; it is also sequence-specific, cleaving proteins on the C-terminal side of lysine residues. There are also less commonly used proteases, such as Asp-N and Glu-C, which, despite being sequence-specific, are less active and thus less efficient. These are sometimes used in combination with a more active protease, like trypsin or Lys-C (Steen & Mann, 2004).

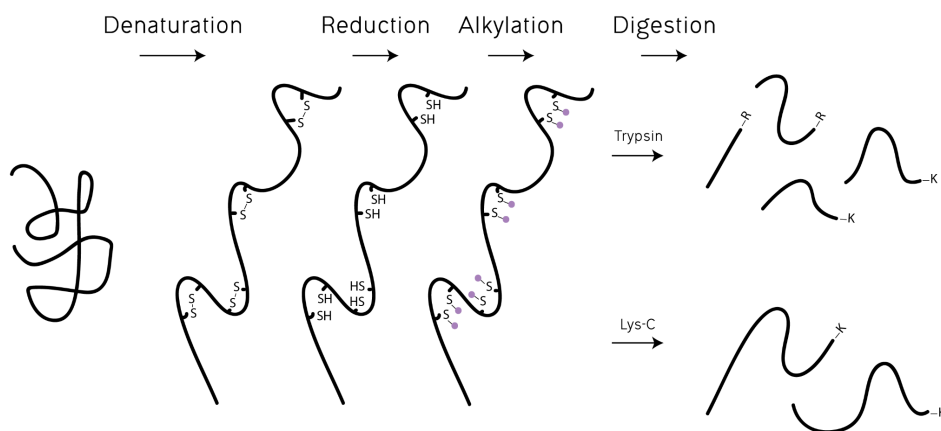


Figure 2. Protein digestion process. Proteins are extracted from a biological sample and denatured with detergents or chaotropic reagents. Disulfide bonds are then reduced with DTT and alkylated with IAA. Finally, proteins are digested with sequence specific proteases, such as Trypsin and Lys-C to obtain different peptide mixtures. DTT, dithiothreitol; IAA, 2-iodoacetamide; R, arginine; K, lysine.

1.2.3 Peptide separation

The peptide mixture resulting from the digestion step is not introduced into the mass spectrometer all at once. Instead, peptides are separated by high performance liquid chromatography (HPLC). Chromatography-based separation relies on the affinity of an analyte, in this case, peptides, to the stationary phase. In bottom up-proteomics, a reversed-phase is usually used as the stationary phase. Its surface is coated with long hydrophobic alkyl chains (C18 material) that better retain hydrophobic compounds than hydrophilic ones. Prior to

separation, the peptides are desalted with stop-and-go-extraction tips (StageTips) using C18 material as the solid phase. Under acidic conditions, positively charged peptides exhibit a high affinity to the C18 material and are thus retained while being washed, until salts are removed (Rappsilber et al., 2007). Then, peptides are eluted using an organic solvent such as acetonitrile (ACN) and loaded into the HPLC column at very high pressures (up to 1,200 bar).

Once in the column, peptides are separated based on their hydrophobicity. A gradient with increasing ACN concentration is set up, causing more hydrophilic peptides to elute first, and more hydrophobic peptides to elute last. Peptides are eluted in as small a volume as possible as the MS signal intensity is proportional to the analyte concentration. This can be achieved by making the chromatographic column with a small inner diameter, usually between 50 and 150 μm (Steen & Mann, 2004). Additional factors determining the HPLC resolving power are the column length, which ranges from 10 to 60 cm, and gradient length, which ranges from a few minutes to a few hours. Longer columns and gradients result in increased peptide identifications (Hsieh et al., 2013; Jorgenson, 2010; Köcher et al., 2011).

1.2.4 Peptide ionisation

Separated peptides, with a high hydrophobicity and a mass in the range of a few kDa, need to be transferred into the gas phase before entering the mass spectrometer. Soft ionisation techniques allow for the gentle ionisation of analytes, in this case, peptides, without causing significant fragmentation of the molecules. Additionally, they are highly sensitive and versatile; they work well with low-abundance molecules, and they are compatible with different solvents and MS detectors. Thus, soft ionisation techniques, such as matrix-assisted laser desorption/ionisation (MALDI) and electrospray ionisation (ESI), are typically selected for peptide ionisation in the context of MS (Steen & Mann, 2004).

1.2.4.1 MALDI

In a MALDI setup (Figure 3a), the previously separated peptides are mixed with an excess of ultraviolet-absorbing matrix. Its main component is typically a low-molecular-weight aromatic acid, such as the gentisic acid or the sinapinic acid. A solution containing the selected compound and an organic solvent, such as ACN, is mixed with the sample's peptides and placed on a metal plate. There, the solvents evaporate, leaving behind a crystalized matrix containing the embedded peptides (Karas & Krüger, 2003).

At this point, the matrix is irradiated with a focused laser beam; the aromatic acid in the matrix absorbs the laser's energy, thus triggering the ablation and desorption of the matrix. The appropriate laser wavelength is connected to the acid's absorption range. Additionally, the process is favoured by the aromatic acid's low molecular weight. The non-volatile peptides, embedded in the matrix, transition into a gas-phase as the excess of aromatic acid sublimates. Finally, a plume of ions and molecules is generated; within this plume, peptides undergo protonation, a process favoured by the proton-donating nature of the aromatic acid compound. This completes the ionisation process, and the ionised peptides are then accelerated by electric potentials into the mass spectrometer (Karas & Krüger, 2003; Karas & Hillenkamp, 1988; Steen & Mann, 2004).

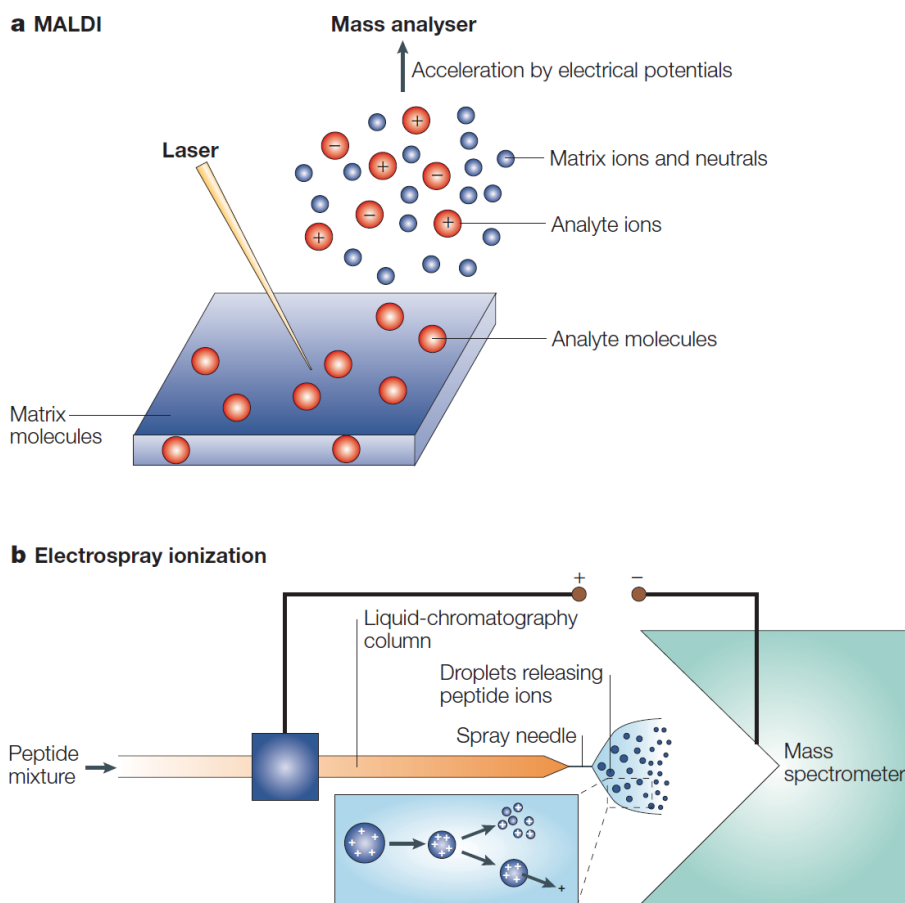


Figure 3. Soft ionisation techniques. MALDI (a) and ESI (b) are two highly sensitive and versatile peptide ionisation techniques often coupled to mass spectrometry setups. Both allow the gentle ionisation of peptides without causing significant fragmentation of the molecules. MALDI, Matrix-assisted laser desorption/ionisation; ESI, electrospray ionisation. From (Steen & Mann, 2004).

1.2.4.2 ESI

In an ESI setup (Figure 3b), the eluted peptides from an HPLC column are subjected to a high electric potential (e.g., 2.4 kV). Using a spray needle, the flow rates are reduced to the nanoliter range (e.g., 200 nL/min), which, under the electric potential, generates charged droplets. These droplets consist of both the solvent (a solution containing an organic solvent, such as ACN) and the analyte (the separated peptides). These highly charged droplets carry a positive charge, resulting from an excess of protons introduced during the peptide digestion step when cleaving after lysines and arginines (Steen & Mann, 2004).

As the solvent is exposed, it starts evaporating, decreasing the droplet size and increasing its charge density. Finally, peptide desolvation is achieved either by their desorption from the droplet surface due to high electrical fields, or by the repetitive fission of droplets, which eventually results in droplets containing, on average, one peptide ion (Steen & Mann, 2004). The development of the ESI technique earned John B. Fenn a shared Nobel prize for chemistry in 2002 (Fenn et al., 1989).

1.2.5 Ion mobility spectrometry

Ion mobility spectrometry (IMS) techniques are based on how ions move in gases in response to exposure to an electric field. Smaller ions, which are more mobile, travel faster in specific electric fields than larger ions, which are less mobile. Thus, IMS relies on either space or time, depending on the set-up, to separate ions based on these mobility differences.

In MS context, IMS adds another layer of peptide separation, this time at ion level. It provides additional analytical power as it enables multidimensional characterisation of detected analytes. Thus, peptides that simultaneously elute from an HPLC column, can be separated at ion level. In this way, IMS enables the separation of isoforms (Dodds & Baker, 2019).

1.2.5.1 TIMS

In trapped ion mobility spectrometry (TIMS), ions are held stationary, hence “trapped”, against an opposing electric field (Figure 4a). After entering an entrance funnel, which controls ion deflection and focusing, ion motion is directed toward the mass spectrometer by gas flow. Before reaching the exit funnel, the TIMS tunnel accumulates, traps, and elutes ions in response to the interplay between the parallel gas flow and the opposing electric field. Thus, the electric field strength is decreased to eject ions at specific mobilities, achieving ion separation in this manner (Dodds & Baker, 2019; Michelmann et al., 2015).

1.2.5.2 FAIMS

In field asymmetric waveform ion mobility spectrometry (FAIMS), a periodic, asymmetric waveform is applied to separate ions (Figure 4b). Ions, carried by a flowing gas, pass through two parallel plates to which particular voltages can be applied. Typically, while the ground plate is held at ground potential, an electric field of oscillating strength is generated, alternating between high and low voltage, on the upper plate. The waveform created on the upper plate modifies the ion trajectory as ions flow between the plates. Additionally, a compensation voltage (CV) is applied. The CV is adjusted to a particular polarity and magnitude that matches a certain range of ions. Ions matching the CV continue to travel between the two plates, exiting the FAIMS device and entering the mass spectrometer. In contrast, ions that do not match the CV collide with either plate and are removed. Thus, FAIMS devices are operated with different CV scans; ions responding to changes in their mobility, triggered by the predetermined CV, are transmitted into the mass spectrometer achieving ion separation (Dodds & Baker, 2019; Kolakowski & Mester, 2007).

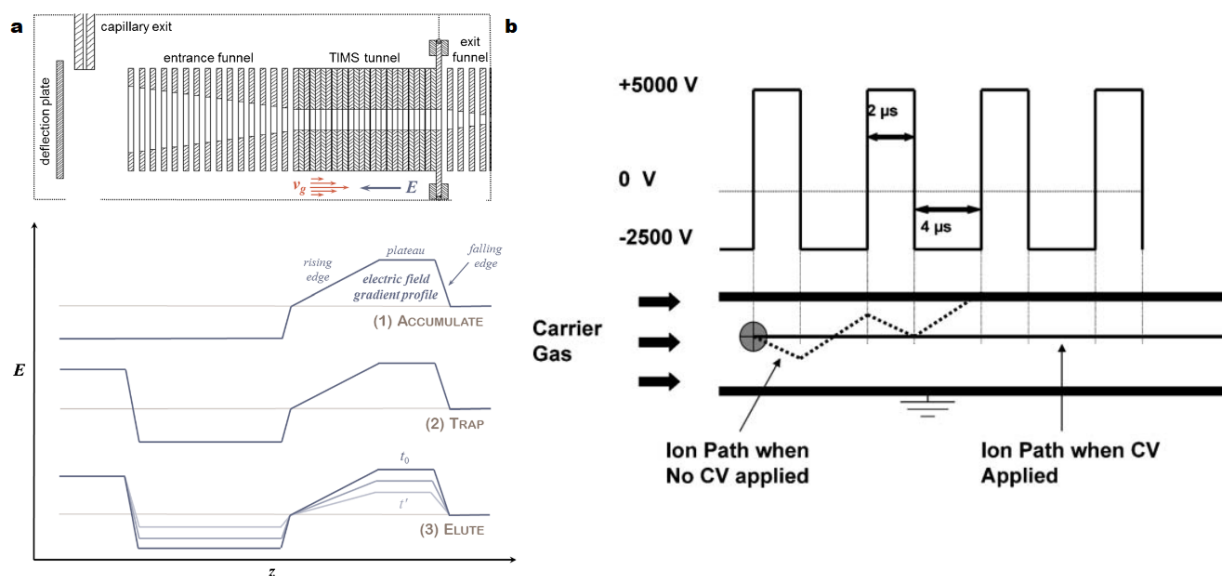


Figure 4. Ion mobility spectrometry techniques. (a) TIMS device diagram, showing the entrance, TIMS and exit funnels. Ions are trapped and accumulated when opposing electric field (E) to the ion's flow (v_g). Variation on the electric field results in the ion's elution. (b) FAIMS device diagram, showing the parallel plates and the voltage asymmetric wave applied to the upper one. The ion's path is dependent on CV. TIMS, trapped ion mobility spectrometry; FAIMS, field asymmetric waveform ion mobility spectrometry; CV, compensation voltage. (a) adapted from (Michelmann et al., 2015) and (b) adapted from (Kolakowski & Mester, 2007)

1.3 Mass spectrometer

There are three main components of a mass spectrometer: (i) the ion source, (ii) the mass analyzer, and (iii) the detector. Mass spectrometers can only analyse gaseous ions. The first component, the ion source, ionises the analytes and transfers them into the mass spectrometer. As discussed previously, MALDI and ESI are two widely used MS ionisation techniques.

The second component, the mass analyser, separates ions based on their movement, which ultimately relies on their m/z ratio. Mass analysers can be grouped into two categories: beam type, such as the quadrupole and the time-of-flight (TOF), and trap type, like the orbitrap. In bottom-up proteomics, tandem MS, where two mass analysers are coupled, is frequently used (Figure 5). Typically, it involves a quadrupole, combined with either a TOF or an orbitrap mass analyser. In tandem MS, two mass spectra are generated: the first one, called MS^1 or precursor ion spectrum, corresponds to the parental peptide ions. These ions then enter a collision cell where they are fragmented. The resulting fragmented ions enter the mass analyser, where the MS^2 or product ion spectrum is generated. Finally, a detector specific to each mass analyser records the ion signals used for peptide identification and quantification (de Hoffmann, 1996).

In addition to the three main components, MS setups also include a combination of peptide and ion separators. These are coupled to the setup before the mass analyser component; peptide separation devices, such as HPLC, are coupled before the ion source, while ion separation devices are coupled after the ionisation process and before entering the mass analyser (Sinha & Mann, 2020).

MS data quality is tightly related to the selected mass analyser, as it determines (i) mass accuracy and resolution, (ii) dynamic range, and (iii) scan speed. (i) Mass accuracy is defined as the difference between measured and theoretical mass, determining in high resolution instruments the mass error in part per million (ppm). Mass resolution describes the capacity of a mass analyser to discriminate between two ions with similar m/z values. Mass resolution also affects whether the mass accuracy is high or low; typically, high-resolution mass analysers achieve a sub-ppm mass accuracy (Makarov & Denisov, 2009; Olsen et al., 2005). (ii) The dynamic range measures the capability of the mass analyser to detect low-abundance ions in the presence of higher abundant ions. Detection not only relies on the detection principle but also on the complexity of the peptide mixture; peptide abundance can vary by several folds depending on the biological source. (iii) Finally, the scan speed parameter, measured in Hz, reflects the number of MS/MS spectra measured in one second (Scigelova & Makarov, 2009).

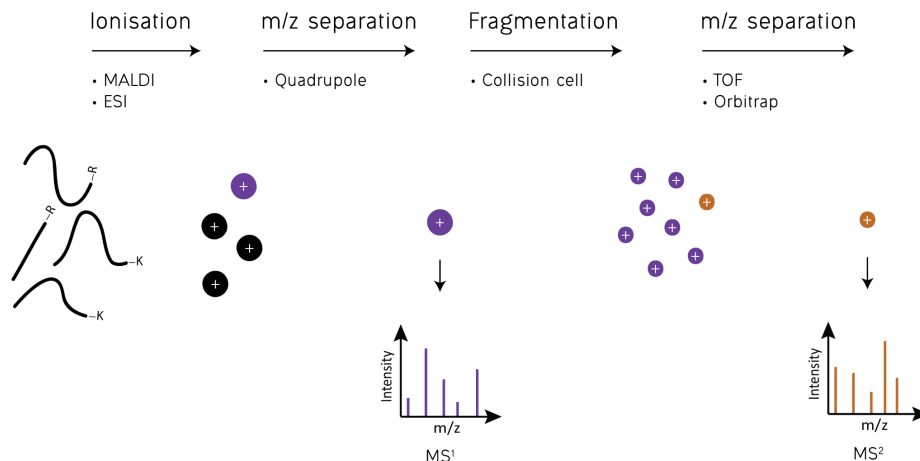


Figure 5. Tandem mass spectrometry. In bottom-up proteomics, digested peptides are identified and quantified using mass spectrometry. After peptide ionisation, ions enter a first mass filter (usually a quadrupole) where they are separated by m/z. Then, the precursor ion spectra, also known as MS¹, are recorded prior to ion fragmentation in a collision cell. Finally, fragmented ions are separated and measured in a mass analyser, such as a TOF or an orbitrap, and the product ion spectra, also known as MS², are recorded R, arginine; K, lysine.

1.3.1 Quadrupole - Time of flight

In a quadrupole-time of flight (QTOF) MS setup (Figure 6a), the first component, the quadrupole, functions as a mass filter. It consists of four cylindrical rods accurately arranged in a parallel radial array, acting as electrodes. Ion separation based on m/z ratio is achieved by applying a combined constant (dc) and variable (ac) electric potential. The dc for each pair of rods is tuned to produce a phase offset; they have equal magnitude but opposite signs. In a typical proteomics measurement, the ion beam, carrying a positive charge, is focused and accelerated towards the quadrupole exit by the positively charged rods. Conversely, the negatively charged rods attract and defocus the beam. Additionally, when the ac variable waveform is applied, the ion beam also undergoes focus/defocus cycles based on the waveform's sign and frequency (Miller & Denton, 1986).

At high frequencies, heavier ions tend to ignore the ac effect. Consequently, they maintain a steady path towards the quadrupole exit because they are primarily influenced by the average quadrupole potential. In contrast, lighter ions are more sensitive to the ac effect, which may lead to their collision with the rods. Therefore, ions with m/z values below a critical threshold are defocused from the beam and filtered out. This configuration allows the quadrupole to selectively isolate ions within a specific m/z range. The signal from these ions is recorded in the MS¹ spectrum and they are transferred to the collision cell (Miller & Denton, 1986).

Typically, in QTOF setups, the collision cell consists of a second quadrupole where ions undergo collision-induced dissociation (CID). Selected ions are accelerated by applying an electric potential, leading to collisions with neutral molecules (such as helium, nitrogen or argon). These collisions result in ion fragmentation. The smaller ion fragments are then also analysed with the mass analyser (Sinha & Mann, 2020).

TOF mass analysers determine the time ions take to travel a known distance. To achieve this, ions are accelerated upwards into a flight tube. The velocity they attain depends on their m/z ratio; heavier ions at the same charge will reach lower velocities. However, ions enter the flight tube with different initial accelerations, causing variations in their velocities, even at the same m/z ratio. The reflectron, located at the upper end of the flight tube, helps reduce this variation. It does so by reversing the direction of ion flight. Ions with higher velocities take a longer path to reverse compared to ions at lower velocities. Consequently, ions with the same m/z ratio require the same amount of time to reach the microchannel plate (MCP) detector, positioned at the lower end of the flying tube. When they reach the MCP, each individual fragmented ion ejects electrons. These electrons are amplified and recorded to generate the MS spectra (Boesl, 2017; Mamyrin, 2001).

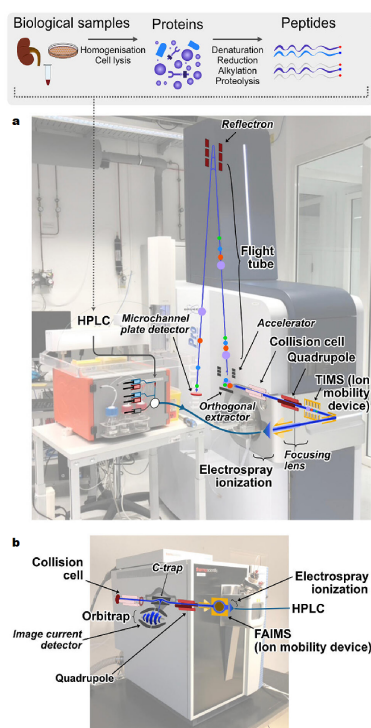


Figure 6. Mass spectrometry setup overview. Proteins are extracted from a biological sample and digested into peptides. These are later separated by hydrophobicity with an HPLC system and ionised by ESI. At this point the ions are further separated and analysed with (a) a TIMS-QTOF mass spectrometer or with (b) a FAIMS-Quadrupole-Orbitrap mass spectrometer. Adapted from (Sinha & Mann, 2020).

1.3.2 Quadrupole - Orbitrap

In a quadrupole-Orbitrap MS setup (Figure 6b), the quadrupole operates based on the same principles as in the QTOF setup. Therefore, its electric potential configuration determines the m/z ratios at which ions are selected. It is typically set to a broad range (300 - 1650) to record spectra for as many parent ions as possible. The selected ions are then subjected to fragmentation through a CID process known as higher-energy C-trap dissociation (HCD). After the quadrupole mass filtering, the ions are temporarily stored into a C-trap and directed into a collision cell. There, high voltages induce collisions with a neutral gas, resulting in ion fragmentation (Olsen et al., 2007). The smaller ion fragments are subsequently sent back to the C-trap before their analysis with the Orbitrap mass analyser.

Orbitrap mass analysers separate fragmented ions based on their oscillation frequencies along a central metal spindle electrode. The central metal spindle electrode is surrounded by two outer electrodes. An electric field is generated by applying voltage between the outer and central electrodes. Fragmented ions are tangentially injected into the Orbitrap, where, under the influence of the electric field, they follow a nearly circular spiral path around the central spindle. This motion creates oscillation frequencies that are dependent on the m/z ratio of each fragmented ion. These oscillation frequencies are captured by an image current detector. The image current, recorded in the time domain, is later transformed into the frequency domain using a Fourier transformation, which in turn generates the MS spectra (Makarov, 2000; Makarov et al., 2006; Zubarev & Makarov, 2013).

1.4 Data acquisition strategies

MS setups are operated with several data acquisition strategies. In bottom-up proteomics, two widely applied strategies are data dependent acquisition (DDA) and data independent acquisition (DIA) (Figure 7). At any given chromatographic retention time, hundreds of peptide ions enter the mass spectrometer, and they are all measured and recorded in the MS^1 spectra. Then, a selection of ions is fragmented and recorded in the MS^2 spectra. DDA and DIA strategies differ in how the fragmented ions are selected (Sinha & Mann, 2020).

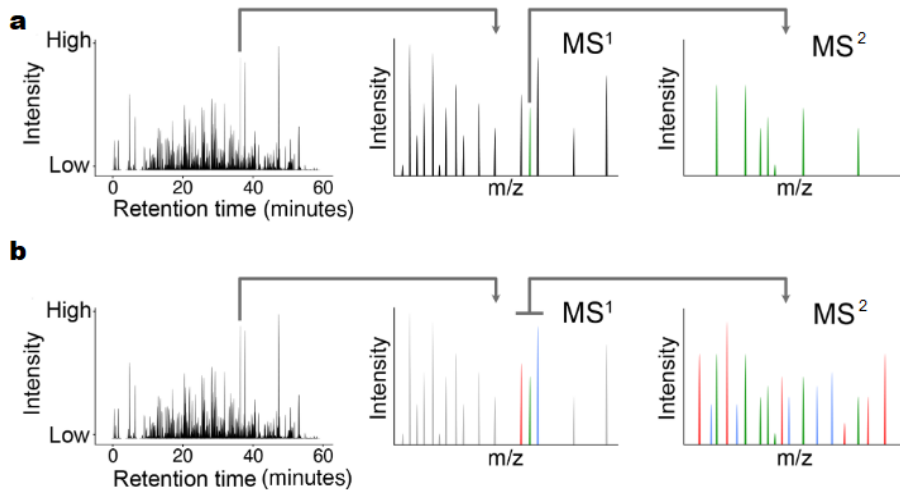


Figure 7. Data acquisition strategies. (a) With the DDA strategy, a singular ion peak is selected from the many available at MS¹ at a particular retention time. This peptide ion is fragmented to obtain the MS² spectra. (b) With the DIA strategy, a range of ion peaks at a particular retention time is chosen. All peptide ions included in the range are fragmented to obtain the MS² spectra. Adapted from (Sinha & Mann, 2020).

Within the DDA strategy (Figure 7a), the user pre-defines a set of rules, including ion's m/z ratio, charge, and intensity, among others. Singular ion peaks are selected according to these rules for fragmentation and MS² spectra measurement. The user also defines the number of selection cycles per MS¹ spectra. It is common to describe DDA acquisition strategies as TopN methods, where N indicates the number of selected peaks, and, consequently, measured MS² spectra. The selection is, this way, partially stochastic because there are more peptide peaks than measurement time. This results in generating missing values for peptides in the sample. However, the advantages of the DDA strategy overcome its limitations. These advantages include ease of setup, data analysis, sample multiplexing and a coverage depth reaching 10,000s of peptides per sample (Aebersold & Mann, 2016; Sinha & Mann, 2020).

Within the DIA strategy (Figure 7b), the parameters configuration focuses on the m/z ratio window selection for the MS² spectra. Instead of selecting particular m/z ratios, a range is chosen. Thus, multiple peptide ions are simultaneously measured at the MS² level. This generates inherently complex spectra containing superimposed fragmentation patterns. Deconvolution of the MS² spectra usually relies on a comparison with previously acquired peptide libraries. This makes the DIA strategy computationally complex, requiring larger computational resources than the DDA strategy. Nevertheless, by covering the full MS¹ m/z ratio range, the DIA strategy generates a robust, unbiased proteome measurement without missing peptide values that also reach 10,000s of peptides per sample (Chapman et al., 2014; Ludwig et al., 2018).

1.5 Mass spectrometry spectra

Tandem-MS results in two kinds of mass spectra: MS¹, or precursor ion spectra, and MS², or product ion spectra. They are both a representation of intensity over m/z ratio (Figure 8). MS¹ covers all peptide ions eluted at a particular time in the HPLC gradient (Figure 8a). In DDA shotgun proteomics, a popular strategy is to select the top-N most abundant peptide peaks for fragmentation; these are determined based on their abundance in the MS¹ spectra. Thus, the peaks observed in the MS² spectra are a product of the peptide ion fragmentation of top-N MS¹ peaks; for each selected peak, an MS² spectra is generated (Figure 8b). When a peptide ion is under the forces generated at the collision cell, it can be fragmented at multiple sites of the amino acid backbone; this generates a different collection of peptide ions that differ, at least, by one amino acid. These are captured at the MS² spectra, which shows the mass differences between the peptide ions in the collection revealing the precursor peptide ion sequence (Steen & Mann, 2004).

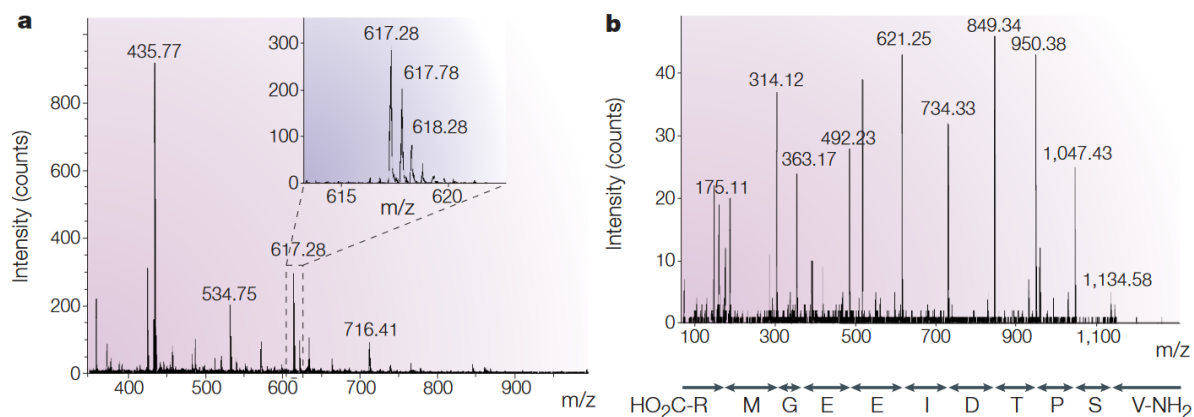


Figure 8. Tandem-MS spectra. MS spectra are a representation of intensity over m/z ratio being (a) the MS¹ spectra, with all peptide ions at a particular HPLC gradient elution time, and (b) the MS² spectra and its associated peptide sequence, resulting from the fragmentation of a selected MS¹ peak (dashed box in a). m/z ratio values are shown at the top of different peptide ion peaks. Adapted from (Steen & Mann, 2004).

A peptide that is fragmented into ions is designated following the Roepstorff-Fohlmann-Biemann nomenclature (Biemann, 1992; Roepstorff & Fohlman, 1984). According to this nomenclature, fragmented peptide ions are labelled consecutively along the peptide backbone. When the charge is retained at the amino-terminal fragment, they are designated as a_m, b_m, or c_m, with m representing the total number of amino acid side chains (Residue, R). The specific designation depends on whether the cleavage occurs at the

α -carbon-carboxyl group bond of the m residue (R_m), the amide bond, or the α -carbon-amino group bond of the $m+1$ residue (R_{m+1}), respectively. Similarly, when the charge is retained at the carboxyl-terminal fragment, ions are labelled z_{n-m} , y_{n-m} , or x_{n-m} , with n representing the total number of R groups and m indicating the number of R groups that the corresponding c_m , b_m or a_m ion would contain. The specific designation depends on the cleavage site and corresponds to their c_m , b_m and a_m counterparts, respectively (Figure 9a) (Steen & Mann, 2004).

In MS setups, peptide ion fragmentation is usually induced with an HCD process in the collision cell. Under these conditions, the fragmentation process has been modelled, and it primarily occurs at the lowest energy bond: the amide bond (Olsen et al., 2007; Zhang, 2004). Thus, in shotgun proteomics, where doubly charged tryptic peptides are commonly found, MS² spectra are dominated by singly charged y_m and b_m ions (Figure 9b).

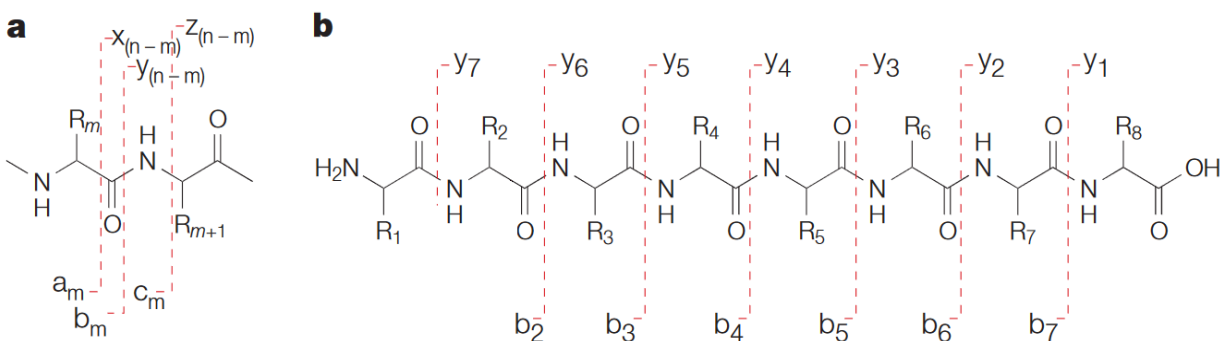


Figure 9. Roepstorff-Fohlmann-Biemann fragmented peptide ion nomenclature. (a) Fragmented peptide ions are labelled, consecutively, along the peptide backbone. The label is designed by determining at which bond the fragment cleaves and whether the charge is retained at the amino or carboxyl terminal group. (b) The amide bond has the lower energy, so b_m and y_m ions dominate CID shotgun proteomics MS² spectra. Adapted from (Steen & Mann, 2004).

1.6 Data analysis

Proteomics data analysis covers several aspects: from the association of the measured peptide ions' MS spectra with protein sequences and their quantification, including exploratory analyses and quality control of the data, to statistically determining differences in protein abundance between samples. The following section will cover the basics steps in proteomics data analysis.

1.6.1 Peptide identification

MS peptide spectra need to be identified and associated with their corresponding proteins. In principle, one can approach a de novo sequencing strategy where the amino acid sequence is determined by considering the mass difference between neighbouring peaks in a collection of fragmented peptide ions (Figure 8b). The success of this approach heavily relies on the quality of the data, which is tightly related with the MS instrument in terms of mass accuracy and resolution, as well as the complexity of the sample (Steen & Mann, 2004).

A far more applicable approach for peptide identification is to use a database-matching strategy. To do so, search engines are configured to uniquely match MS spectra-detected peptides to known peptide sequences. The success in this strategy relies on the fact that an organism's protein set is defined by its genome sequence. Hence, the combination of amino acid sequences one can find in a given proteome is known and finite. Thus, for a given organism, its proteome amino acid sequences can be stored in a database. These sequences are then *in silico* digested at a selected protease cleavage site and fragmented to generate theoretical spectra to match them to the observed spectra. Thus, while an MS² spectra might not contain the whole information to determine a peptide's full amino acid sequence, it might contain enough data to match with a high statistical significance to a unique peptide sequence derived from the *in silico* digestion (Allet et al., 2004; Taylor & Johnson, 1997).

A popular peptide identification search engine is the Andromeda search engine as implemented in the MaxQuant software (Cox et al., 2011; Cox & Mann, 2008). The MaxQuant search engine relies on the match between observed and expected spectrum. For each match, the Andromeda algorithm calculates a probability score of peptide-spectrum matches (PSMs). Then, rather than deciding individually whether a PSM is correct or incorrect, Andromeda follows a target-decoy database implementation. Within this strategy, a composite database is created, including both a target protein sequence database and a decoy database. The target protein sequence database is selected according to the analysed protein mixture, while the decoy database contains the reversed sequences in the target database. The reversed sequences are obtained after *in silico* digestion and maintain the last amino acid unreversed so the protease cleavage point is maintained. The target-decoy strategy allows the evaluation of false positives (FP) rates through large PSM populations using an FP rate across the entire set. Ultimately, this strategy provides statistical evidence on the quality of each PSM and limits the overall number of incorrect identifications (Elias & Gygi, 2007). Another feature included in the Andromeda search engine is a second peptide search for each MS² spectrum, which enables the identification of co-fragmentation signals in the same MS² spectrum resulting from additional

co-eluting precursor ions (Cox et al., 2011). Besides MaxQuant and the Andromeda search engine, other popular search engines, such as Mascot (Perkins et al., 1999) and MSFragger (Kong et al., 2017) search engines are available. Mascot operates with a similar strategy and performance than Andromeda. On the other hand, MSFragger relies on a fragment-ion indexing strategy, which results in faster protein identifications and an overall shorter computational time when compared to most search engines, especially when configured for detecting protein modification profiles (Kong et al., 2017).

1.6.2 Quantification strategies

There are several approaches to obtaining quantitative proteomics data using MS techniques. They all rely on comparing peptide signals from the MS spectra under different conditions to determine protein abundance or estimate it. Like any other type of quantitative data, protein abundance can be obtained as an absolute measurement (the amount of protein in a particular sample) or as a relative measurement (the amount of protein in relation to another measurement of the same protein in a different sample) (Ong & Mann, 2005).

1.6.2.1 Label-based quantification strategies

Label-based quantification is closely related to the use of stable isotopes with a low percentage of natural abundance, such as hydrogen (^2H), carbon (^{13}C) or nitrogen (^{15}N). These isotopes are introduced into a protein mixture, resulting in a mass shift that can be detected with MS. Comparing samples with and without isotope labelling allows protein quantification. This is nowadays usually accomplished by comparing the area under the curve from the extracted ion current (XIC), in a process known as XIC-based quantification. MS spectra, before undergoing Fourier transformation, are expressed as peptide signal intensity over HPLC gradient elution time. Under the same experimental conditions, the area under the curve in such plots for a particular peptide is proportionally related to its abundance. In label-based quantification strategies, the XIC of labelled and unlabelled peptides share the exact same experimental conditions; they are detected on a single MS chromatogram. This leads to accurate protein abundance ratio determination. Thus, XIC is the standard quantification method for label-based techniques (Ong & Mann, 2005). These techniques are then broadly classified as chemical or metabolic labelling, depending on how the stable isotopes are introduced on the protein mixture (Figure 10) (Bantscheff et al., 2007).

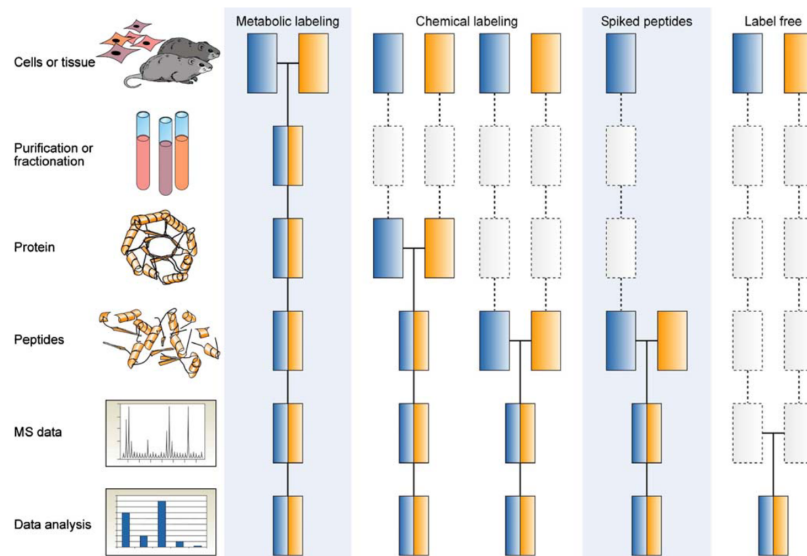


Figure 10. MS quantification strategies. The most common quantification strategies are schematically represented. Blue and yellow boxes each represent an experimental condition, while dashed boxes indicate sources of experimental variation. The later the samples are combined, indicated with an horizontal line, the greater is the experimental variation enabling for greater quantification errors. From (Bantscheff et al., 2007).

Chemical labelling can be applied either at protein or at peptide level, depending on each particular technique; this leaves room for variability and potentially lower robustness, especially if the label incorporation is not homogeneous throughout the samples. Nevertheless, there are several popular chemical labelling methodologies that result in fast and inexpensive workflows (Bantscheff et al., 2007). Chemical labelling relies on the modification of either reactive sites on amino acids side chains or protein terminal groups, if not both. For instance, within the dimethyl labelling technique, primary amines in peptide sequences (N-terminal group and lysine ϵ -amino group) are converted to dimethylamines through reductive amination. The reaction is triggered by formaldehyde (in the presence of NaBH_3CN) and it introduces a +28 Da mass shift. Additionally, isotope-labelled formaldehyde, either with ^2H or with ^2H and ^{13}C , is also used to introduce a +32 Da or +36 Da mass shift, respectively. In both cases, the reaction is triggered in the presence of deuterated NaBH_3CN (Boersema et al., 2009; Hsu et al., 2003). These modifications are later detected through XIC-based quantification, leading to accurate protein ratio determination. Another popular chemical labelling strategy relies on incorporating isobaric tags to amino acid reactive groups; these isobaric tags have identical masses despite having a different distribution of heavy isotopes in their structure. At MS^1 level, isobaric labelled peptides have identical mass, but at MS^2 level, upon fragmentation, distinct reporter ion masses are detected. Tandem mass tag (TMT) (Li et al., 2020; Thompson et al., 2003) and isobaric tag

for relative and absolute quantification (iTRAQ) (Ross et al., 2004) are two techniques relying on isobaric tags. The main advantages of isobaric tag labelling are its high multiplexing and its ease of application to any sample type.

Alternatively, stable isotope labelling can be approached in a metabolic way. In this case, the heavy isotopes are directly incorporated into the newly synthesised proteins by an organism's metabolic pathways, resulting in a defined mass shift. A main advantage of metabolic labelling is its robustness; since samples from both experimental conditions are combined right after protein purification, any variation on the workflow will affect all peptides and conditions equally (Ong & Mann, 2006). One of the most widely used metabolic labelling implementations is the technique known as stable isotope labelling by amino acids in cell culture (SILAC). During a SILAC workflow, cells are either grown with a regular medium or with a medium containing an isotopically labelled (with ^{13}C and/or ^{15}N) analogue of lysine and arginine. When digested with trypsin, all peptides from cell extracts grown with the labelled medium will contain at least one labelled amino acid. This creates a mass shift between labelled and unlabelled peptides which is detected at MS¹ level. Thus, relative quantification is reached by comparing the signal intensities of isotope clusters of labelled/unlabelled peptide pairs with intensity-based quantification (Ong et al., 2002).

Finally, stable isotopes are also used by directly spiking in labelled standard peptides (Figure 10). This enables, for example, an alternative method of quantification called absolute quantification (AQUA). By spiking in selected chemically synthesised labelled peptides during the protein digestion, at a known concentration, one can determine the absolute amount of a specific protein in a sample. This quantification method is popular among targeted proteomics approaches (Gerber et al., 2003; Kirkpatrick et al., 2005).

1.6.2.2 Label-free quantification strategies

Label-free quantification (LFQ) encompasses a range of techniques that share the common feature of not relying on heavy isotopes. LFQ techniques are widely used in shotgun proteomics, and among their main advantages are their straightforward workflows, cost-effectiveness, and independence from the number of samples (Al Shweiki et al., 2017; Bantscheff et al., 2007).

Among the most basic forms of LFQ, we find the protein abundance index (PAI) and its exponentially modified (emPAI) version. PAI relies on peptide counting and is calculated by dividing the measured peptide count associated with a protein by the theoretical number of peptides obtained after digesting the protein (Ishihama et al., 2005; Shinoda et al., 2010).

XIC-based quantification can also be used for LFQ. As long as a peptide is measured, one can compare two different chromatograms to extract relative quantification ratios. To obtain accurate results using an LFQ XIC-based approach, one should carefully monitor the sample processing, as this technique is prone to quantification errors due to experimental variations in the workflow (Ong & Mann, 2005). Another LFQ technique, used to estimate absolute protein abundances, is intensity-based absolute quantification (iBAQ). During iBAQ, all peptide intensities matching a protein are summed and then divided by the number of theoretical peptides, resulting in an approximation of protein abundance (Schwanhäusser et al., 2011).

A more sophisticated LFQ algorithm is MaxLFQ, implemented in the MaxQuant software. MaxLFQ includes some features that enhance its quantification accuracy: (i) MaxLFQ allows for the transfer of peptide identification between different MS samples with its “match between runs” option. A peptide might be detected at MS¹ level but not selected for fragmentation, thereby missing its MS² spectrum. If the same peptide is detected at MS¹ level in a different sample from the same experimental set, activating the “match between runs” option will result in the peptide being identified in both samples. (ii) MaxLFQ controls for experimental variability between samples by normalising peptide intensities and calculating associated protein intensities using multiple pairwise peptide ratios. (iii) MaxLFQ can handle missing peptide intensity values due to the dynamic range of protein abundances or technical limitations. It employs a probabilistic framework to impute missing values based on neighbouring peptides, enabling robust quantification even in complex datasets. (iv) MaxLFQ also accounts for low-abundance proteins by incorporating intensity information from all identified peptides, including those with low signal-to-noise ratios. This enhances the sensitivity of protein detection. As a result, MaxLFQ generates an LFQ intensity estimate for each identified protein in a sample (Cox et al., 2014; Cox & Mann, 2008).

1.6.3 Missing value imputation

In DDA proteomics, missing values in the data are highly frequent, especially when using LFQ. Missing values are broadly classified into two categories. On one hand, there are missing completely at random (MCAR) and missing at random (MAR) values. These account for peptides that were not measured due to technical reasons, such as miss-cleaved peptides, unsuccessfully ionised peptides, and other biochemical, analytical and computational issues. MCAR and MAR values should evenly affect the entire data set (Karpievitch et al., 2012; Lazar et al., 2016).

On the other hand, there are missing not at random (MNAR) values. These account for peptides that were not measured either due to biological reasons (e.g., protein expression suppression in one of the measured conditions) or due to low peptide abundance, which is too close to the mass spectrometer's limit of detection. MNAR values have a targeted effect, with low-abundance peptides having a higher rate of MNAR values (Lazar et al., 2016).

Data with missing values hinders downstream analyses, such as the statistical assessment of protein abundance differences. Hence, several imputation methodologies have been proposed to fill in missing values. For instance, the probabilistic minimum imputation (MinProb) method replaces the missing values with random draws of a Gaussian distribution centred on the lowest value detected in the data set (Chich et al., 2007). The MinProb approach focuses on left-censored data, which in a proteomics data set corresponds to the region where low-abundance peptides fall. Thus, it deals effectively with MNAR values and has gained widespread usage in the proteomics community. Nevertheless, MinProb is just one among several imputation methods; one should consider which kind of missing values are more abundant in the data set before selecting an imputation methodology (Lazar et al., 2016).

1.6.4 Exploratory data analysis

Exploratory data analysis encompasses a number of statistical techniques aimed at describing and summarising the characteristics of a data set. Their applications are universal and, as result, are regularly used in proteomics data analysis workflows. The main goals of exploratory analysis are to assess data quality and help formulate hypotheses to be statistically tested (Tukey, 1977).

A standard proteomic data set can be described as a matrix where rows are quantified proteins, and columns are different MS-measured samples. Each cell in the matrix contains the protein abundance value, which is dependent on the quantification technique. One of the characteristics one can explore from such a matrix is how similar (or dissimilar) its columns (MS samples) are. Sample similarity serves as a quality control measurement. In experimental set ups where multiple replicas of the same condition were measured, a high degree of sample similarity among replicas generally indicates experimental workflow reproducibility. Conversely, dissimilarities between samples from different conditions might help formulate hypotheses that could explain such differences.

One of the most common ways to measure the similarity between samples is to calculate their Euclidean distance. This distance is calculated by subtracting, for each protein, the abundance between two samples; then, each difference result is squared. Finally, squared

differences are summed and square-rooted, removing any negative value. The Euclidean distance provides a straightforward and easy to interpret measurement of sample similarity. Smaller distances indicate higher similarity, while larger distances indicate greater dissimilarity. Another popular measurement on sample similarity is the Pearson's correlation coefficient. It is calculated from the covariance between two variables divided by the product of their standard deviations. In proteomics, the two variables are represented by protein abundances for each pair of samples. Pearson's correlation coefficients range from -1 to 1. The closer a coefficient gets to 0, the less correlated two samples are and, therefore, the more dissimilar they are. A score of 1 indicates a perfect linear correlation and similarity (e.g., when comparing a sample to itself), while a score closer to -1 indicates a negative correlation between samples.

Euclidean distance and Pearson's correlation coefficient are often combined with sample unsupervised hierarchical clustering, another exploratory data analysis technique. In the agglomerative hierarchical clustering approach, each sample starts as an individual cluster, and then, through iterative merging, the most similar clusters are combined until a single cluster encompassing all samples is obtained. As the merging process continues, a dendrogram is constructed—a tree-like diagram that illustrates the hierarchical relationships between clusters. The height of the branches in the dendrogram represents the dissimilarity between clusters or samples. Hierarchical clustering dendrograms and sample similarity measurements, such as Euclidean distance and Pearson's correlation coefficient, are often combined into heat maps, which offer an intuitive data representation, enhancing its interpretability. In addition to hierarchical clustering, other unsupervised clustering methods applied to proteomics data sets include k-means clustering and the self-organising map (SOM) machine learning algorithm.

Another popular exploratory data analysis strategy involves data dimensionality reduction. The main goal of such strategies is to reduce the number of variables in a dataset while preserving the maximum amount of information, enhancing data interpretability, and revealing underlying patterns and sources of data variation, providing insights into the structure and relationship between MS samples. Principal component analysis (PCA) is a widely used technique for feature dimensionality reduction. In a proteomics data set, each measured sample represents a dimension, with abundance values for each protein and sample. Principal components describe the variation within these dimensions, resulting in one principal component per sample. PCA assigns a value to each protein describing its influence on the principal component. When considering a singular dimension, protein abundance is usually expressed as a uniform distribution ranging from lower to higher values. Its principal component would describe the variation within this distribution, and proteins at the distribution edges would

have higher influence values, with opposing signs. The product between all protein abundances and their associated influence is calculated, and then, a principal component score is obtained as the sum of these products. In multi-dimensional scenarios, each dimension yields a principal component and its corresponding score. For instance, a MS experiment with four samples would result in four principal components and 12 scores (one per principal component and sample). PCA scores are used to assess sample similarity, as they reflect variance in protein abundance. Samples with similar abundance profiles exhibit similar PCA scores. Additionally, each principal component has an associated eigenvalue representing the amount of variation it explains, expressed in percentile. Principal components are ordered by their percentile; the first component captures the most variation in the data, the second component the second most variation in the data, and so on. In high-dimensional data sets, only a subset of principal components, selected based on the explained variance percentile, are considered. The selected subset of principal components is often represented as scatter plots in pairs. This allows PCA to identify key sources of variation in proteomics data, revealing hidden data structures, influential proteins, and abundance patterns associated with specific experimental conditions.

1.6.5 Statistical difference assessment

Statistical difference assessment encompasses a number of techniques focused on hypothesis testing. Their goal is to determine if there is a significant difference between the means of two independent groups. Therefore, for a proteomics data set, the null hypothesis is that there are no significant differences between the mean protein abundance of the two groups, while the alternative hypothesis suggests that there is a significant difference. Thus, in an experimental setup where at least three replicas were measured for each condition, the mean protein abundance between conditions can be tested.

The Welch's t-test is a popular approach to assess such hypotheses. It is an alternative to the traditional Student's t-test, used when the assumptions of equal variances and/or sample sizes are violated. One main reason behind Welch's test popularity is its higher versatility compared to the classic Student's t-test. Thus, Welch's t-test, designed for unequal populations, exhibits robust performance even when applied to populations with equal variances. On the other hand, the classic Student's t-test performance suffers when the assumption of equal variance in the population is violated. Nevertheless, Welch's t-test is still restrictive in its assumption of normal sample distribution and observation independence.

Among the test results, a p-value is obtained. This helps either accept or reject the null hypothesis which, in proteomics, claims an equal protein abundance means between two

experimental conditions. A p-value offers a straightforward interpretation: the null hypothesis is rejected when a p-value is lower than a selected threshold value. Common threshold values for statistical significance are 0.05 and 0.01. Thus, the abundance means of a tested protein for two conditions, whose p-value is below the selected threshold, are considered to be statistically different.

Another popular approach is the analysis of variance (ANOVA), which is used to compare the means of more than two groups and allows for assessing protein abundance differences between more than two conditions. Similarly to the classic Student's t-test, ANOVA assumes independence of observations, normal distribution, and homoscedastic variance. ANOVA relies on partitioning the total variation in the data into two components. This way, it differentiates between variation between groups and variation within groups. These two sources of variation are compared to determine whether the observed differences are significant or not.

Statistical difference assessment allows to identify statistically different protein abundance patterns between tested conditions. This information can provide insights into the biological processes and pathways that are affected by the experimental conditions. However, it is important to consider the assumptions and limitations of the chosen statistical test and interpret the results in conjunction with other relevant biological information.

1.6.6 Functional analysis

Functional analysis of proteomics data involves examining the biological functions and pathways associated with statistically different proteins to gain insights into the underlying biological mechanisms and interpret the results in a biological context. Functional analysis helps understand the functional implications of changes in protein abundance profiles and identify key biological processes affected by experimental conditions.

Proteins are annotated using databases containing biologically relevant information. These databases encompass information such as structural protein domains or complete biochemical pathways. Protein families (Pfam) (Mistry et al., 2021), gene ontology (GO) (The Gene Ontology Consortium et al., 2021) or the Kyoto encyclopaedia of genes and genomes (KEGG) (Kanehisa & Goto, 1999) databases are some examples of popular resources for protein annotation. After protein annotation, an enrichment analysis is performed, usually with a Fisher's exact test, to determine whether specific annotations are significantly overrepresented among the differentially abundant proteins. This analysis helps assess whether there is an enrichment of proteins with specific domains, functions, or pathway involvement compared to what would be expected by chance.

1.7 Applications to study RNA-protein interactions

MS-based proteomics techniques have a broad range of applications in molecular biology. For instance, they are used to study proteomes under specific conditions, enabling a global analysis of protein composition for a given sample. Consequently, protein expression dynamics are revealed when comparing multiple samples under different conditions. Examples include protein dynamics across an organism's developmental stages, protein abundance comparisons between treated and untreated samples, or proteome changes upon gene knockout or overexpression (Aebersold & Mann, 2016; de Godoy et al., 2008). Despite its versatility, the study of protein expression dynamics does not excel at revealing RNA-protein interactions. As part of the interactomics field, the study of RNA-protein interactions aims to understand interaction networks and their functional implications, providing a holistic view of biological systems. To achieve this, it relies on mapping and analysing interactions between biomolecules, gaining insights into the underlying mechanisms of cellular processes. Therefore, a better-suited approach to investigate RNA-protein interactions is using specific baits for affinity purification (AP) to capture the aforementioned interaction partners (Bludau & Aebersold, 2020).

AP-MS techniques rely on immobilising a bait of interest on a matrix or a bead to capture its interacting proteins, found in the sample's lysate (Dunham et al., 2012). One approach is to use nucleic acids as baits. For instance, the RNA interactome capture (RIC) technique combines ultraviolet (UV) protein cross-linking to RNA with oligo(dT) capture to identify RBPs. After polyadenylated messenger RNAs (mRNA) bind to the oligo(dT) beads, mRNAs are digested, and UV-crosslinked RBPs are identified and quantified with MS (Baltz et al., 2012; Castello et al., 2012, 2013). Although this technique cannot identify proteins that interact with premature, unspliced, and non-polyadenylated RNA, it is a widespread technique to identify RBPs and has been used in several organisms, such as *S. cerevisiae* (Beckmann et al., 2015; Matia-González et al., 2015; Mitchell et al., 2013), *C. elegans* (Matia-González et al., 2015) or *H. sapiens* (Beckmann et al., 2015). RNA oligos are also used in a targeted way to identify interacting proteins. Specific RNA sequences are used as bait for RBPs, which are later identified and quantified with MS. To do so, results are compared to those obtained from a scrambled control sequence. By doing so, background proteins binding to either the control or the specific sequence are identified and differentiated from sequence-specific interactions, which are found to be enriched (Butter et al., 2009; Scheibe et al., 2012). Moreover, RNA structural motifs are also used, in a similar fashion as RNA sequences, to identify structure specific interactions by MS (Casas-Vila et al., 2020).

Another popular AP-MS approach is to immobilise protein baits with the use of antibodies (Immunoprecipitation-MS, IP-MS), either against the protein itself or a tag fused into it, in order to identify their interacting prey (Smits & Vermeulen, 2016). Unspecific binders are removed via bead washing, and the remaining bound proteins are identified and quantified with MS. Specific bait interactors are then determined by comparing quantified proteins captured by the bait of interest to quantified proteins captured in a control condition. Thus, while the abundances of background proteins are roughly equal in both situations, the abundances of bait-specific interactors are found to be significantly different from the control (Keilhauer et al., 2015; Vermeulen et al., 2008).

These myriad of techniques are applied to a wide range of RNA types, broadly classified into non-coding and coding RNA. On one hand, non-coding RNA refers to a vast and diverse group of RNA molecules that do not code for proteins. Non-coding RNA molecules can be classified into several categories based on their size and function. For instance, transfer RNA (tRNA) and ribosomal RNA (rRNA) are involved in protein synthesis; tRNA molecules facilitate the translation of mRNA into proteins by carrying specific amino acids to the ribosomes, while rRNA is a crucial component of ribosomes, ensuring their structural integrity. Other examples of non-coding RNAs include small nuclear RNA (snRNA), microRNAs (miRNA), and long non-coding RNA (lncRNA) molecules. Their functions are diverse; snRNAs are involved in the splicing of pre-mRNA, while miRNA are regulators of gene expression. Finally, lncRNAs participate in a wide range of cellular processes, including chromatin remodelling, transcriptional regulation among others. On the other hand, coding RNAs provide the template for protein synthesis, enabling the expression of genetic information. In this category we find mRNA molecules. Throughout their life cycle, mRNA molecules interact with several RBPs which determine their cellular fate.

The articles presented in this thesis showcase two applied cases of using MS for investigating RNA-protein interactions. Article I exemplifies how SILAC-based MS proteomics is used to delve into the molecular function of non-coding RNAs. In particular, it focuses on the identifying interaction partners of the lncRNA TERRA. On the other hand, Article II exemplifies how IP-MS-based interactomics is used to investigate RNA-related molecular cell processes from a systems perspective. This way, the mRNA molecule's life cycle and the role RBPs play in it is investigated.

1.7.1 Telomeric repeat-containing RNA origin and interacting partners

TERRA are lncRNA molecules produced from the transcription of telomeric repeats. TERRA transcripts were first detected in *T. brucei* (Rudenko & Van Der Ploeg, 1989). Since then, they have been detected in several organisms, including *H. sapiens* (Azzalin et al., 2007; Schoeftner & Blasco, 2008) and *M. musculus* (Schoeftner & Blasco, 2008). Thanks to their ability to form RNA:DNA hybrids and interact with several protein partners, TERRA molecules serve a wide range of functionalities. These include heterochromatin regulation, telomeric loop formation and telomerase recruitment (Cusanelli & Chartrand, 2015).

In addition to their transcription from telomeres, *M. musculus* TERRA transcripts' genomic origin has also been traced to intrachromosomal repeats. For instance, *M. musculus* TERRA molecules exhibit different colocalization patterns than *H. sapiens* TERRA when investigated with RNA fluorescence in situ hybridisation (FISH). While TERRA RNA-FISH patterns for *H. sapiens* are detected at telomeres, in *M. musculus*, colocalization patterns are mostly found at pseudoautosomal region (PAR), the subtelomeric q end of X/Y chromosomes and, to a much lesser extent at telomeres. Therefore, in *M. musculus*, a distinction exists between TERRA (originating at telomeres) and PAR-TERRA (originating at Xq/Yq subtelomeric regions) molecules (Azzalin et al., 2007; Chu, Froberg, et al., 2017; De Silanes et al., 2014). These differences between *H. sapiens* and *M. musculus* TERRA in terms of genomic origin suggest variations in TERRA localization and function between these species. Hence, further studies on its genomic origin and conserved interaction partners are desirable.

1.7.2 RNA binding protein network-based function assignment

A primary focus when studying RNA-protein interactions is on RBPs. Their binding to mRNAs forms transient ribonucleoprotein (RNP) complexes that determine the downstream effects of the bound mRNAs (Dreyfuss et al., 2002). These downstream effects resonate throughout the entire mRNA life cycle, from its initial processing (Neve et al., 2017; Ramanathan et al., 2016; Wilkinson et al., 2020) and export from the nucleus (Tutucci & Stutz, 2011) to its final degradation in the cytoplasm (Hasan et al., 2014). RNP complexes play a crucial role in regulating mRNA cellular fate within a large interconnected network (Licatalosi & Darnell, 2010). This versatility is facilitated by interactions with unique combinations of RBPs, which can function either as core or regulatory factors (Rissland, 2017). Thus, the combination of RBPs bound to an mRNA determines its cellular fate.

Several large-scale methodologies have been implemented to discover RBPs, resulting in a large increase in the number of proteins being described as RBPs (Hentze et al., 2018). However, to gain a better understanding of their roles in the mRNA life cycle, the expansion of the RBP catalogue must be accompanied by a functional characterization (Kilchert et al., 2020). RBP protein-protein interaction (PPI) partners are obtained with AP-MS techniques, and the interconnectedness of RBP binding to specific subsets of RNAs is used to infer functionality (Hogan et al., 2008). Additionally, to specifically reveal RNA-dependent interactions (RDI), immunoprecipitation beads are treated with RNase A. This treatment results in the loss of all prey that interacts with an RBP bait RNA due to RNA digestion. Consequently, when comparing treated and untreated samples, RDIs are found to be significantly enriched. PPIs and RDIs, along with concurrent interactions with other RBPs, are then used to suggest the involvement of the RBP bait in functional pathways (Klass et al., 2013). With sufficient data, function-based networks of PPIs and RDIs can be created to identify functional associations for previously uncharacterized RBPs.

2. Aims of the thesis

I explored two fields involving RNA-protein interactions using MS-based quantitative proteomics: *M. musculus* and *H. sapiens* TERRA-associated proteins and *S. cerevisiae* network-based function assignment of RBPs.

Telomeres are nucleoprotein structures that protect the ends of eukaryotic chromosomes. They contain active transcription sites that produce the lncRNA TERRA molecules. Additionally, TERRA molecules are also transcribed from intrachromosomal telomeric repeats. TERRA molecules interact with RBP, and together, they determine TERRA functionality. The aim of this project was to further interrogate the genomic origins of TERRA species in *M. musculus* and to investigate their interacting partners. To achieve this, among other techniques, an MS-based quantitative proteomics approach was used. A SILAC TERRA interactome screen was performed to assess the conservation between *M. musculus* and *H. sapiens* TERRA interacting partners.

The number of proteins identified as RBPs has largely increased over the last decade. However, this increase in the number of novel RBPs has not been accompanied by functional characterization. RBPs play critical roles throughout the mRNA life cycle, including its initial processing and export from the nucleus, as well as its transport, localization, translation, and degradation in the cytoplasm. The goal of this project was to identify functional associations for previously uncharacterized RBPs and integrate them into an interaction network. To accomplish this, an interactome immunoprecipitation screen was conducted in *S. cerevisiae* to identify PPIs and RDIs of 40 RBPs involved in different RNA pathways. Novel functional associations were described by identifying concurrently binding RBPs in function-based networks.

3. Articles

3.1. PAR-TERRA is the main contributor to telomeric repeat-containing RNA transcripts in normal and cancer mouse cells

3.1.1 Summary

This project focused on TERRA molecules in *M. musculus*: it explored their genomic origins and compared their interactomes to the *H. sapiens* counterpart.

Differences in behaviour between *M. musculus* and *H. sapiens* TERRA molecules were determined with RNA-FISH. While only a few small *M. musculus* TERRA foci are found at telomeres, *H. sapiens* TERRA foci are recurrently found at telomeres. Additionally, TERRA-FISH intensity in *M. musculus* does not correlate with telomere length, while in *H. sapiens*, the intensity does correlate. Hence, for *M. musculus* TERRA molecules, a different genomic origin outside telomeres was hypothesised. To find putative regions that might transcribe to TERRA molecules, the *M. musculus* genome was scanned for the presence of intrachromosomal telomeric TTAGGG repeats. Then, high-coverage RNA-Seq data was overlapped with such regions to map where TERRA reads are located. This resulted in four major chromosomal regions: known Telo 18q, PAR-Xq/Yq and ChrX *Tsix* locus regions, and a novel Chr2 region. Further analyses with reverse-transcription real-time PCR (RT-qPCR) revealed the PAR region as the major source of TERRA molecules.

Conservation of TERRA-associated functions was evaluated with a SILAC-based AP-MS approach. Here, a biotinylated oligonucleotide with a TERRA sequence was used as bait to identify *M. musculus* TERRA-interacting proteins. To determine the enriched TERRA-interacting proteins, identified and quantified proteins were compared to those obtained using a scrambled control sequence. Functional analysis of the enriched proteins revealed an involvement of TERRA-interacting proteins in RNA metabolism, DNA replication, mitosis and chromatin organisation. Further analysis involved a comparison with publicly available *H. sapiens* TERRA-interacting protein data sets, which revealed that despite having a distinct genomic origin, functions are conserved between *M. musculus* and *H. sapiens*.

In summary, this project explored the distinct origin of non-coding RNA on the example of *M. musculus* TERRA molecules and highlighted the function-conservation between *H. sapiens* and *M. musculus* TERRA-interacting proteins.

3.1.2 Zusammenfassung

In diesem Projekt wurden TERRA-Moleküle in *M. musculus* Zellen untersucht, dabei wurden ihre genomischen Ursprünge erforscht und ihr Interaktom mit den entsprechenden *H. sapiens* TERRA-Molekülen verglichen.

Mithilfe von RNA-FISH wurde ein unterschiedliches Verhalten der TERRA-Moleküle von *M. musculus* und *H. sapiens* festgestellt. Während nur wenige kleine TERRA-Anreicherungen in *M. musculus* an den Telomeren gefunden wurden, reichert sich TERRA bei *H. sapiens* besonders an den Telomeren an. Darüber hinaus hängt die Intensität von TERRA-FISH in *M. musculus* nicht mit der Telomerlänge zusammen, während sie bei *H. sapiens* miteinander korrelieren. Daher wurde in *M. musculus* eine andere genomische Herkunft als die Telomere für die TERRA-Moleküle vermutet. Um potenzielle Regionen zu finden, die zu TERRA-Molekülen transkribiert werden könnten, wurde das *M. musculus* Genom auf das Vorkommen intrachromosomaler TTAGGG Wiederholungen untersucht. Anschließend wurden hochauflösende RNA-Seq-Daten mit diesen Regionen überlappt, um die TERRA Sequenzen zu lokalisieren. Daraus resultierten vier chromosomale Regionen: die bekannten Telo 18q, PAR-Xq/Yq und ChrX-Tsix-Lokusregionen sowie eine neue Chr2-Region. Weitere quantitative Analysen mit reverser Transkriptions-Echtzeit-PCR (RT-qPCR) zeigten, dass die PAR-Region Hauptquelle für TERRA-Moleküle in der Maus ist.

Die Konservierung von TERRA-assoziierten Funktionen wurde mit einem SILAC-basierten AP-MS-Ansatz evaluiert. Dazu wurde ein biotinyliertes Oligonukleotid mit einer TERRA Sequenz als Köder verwendet, um interagierende Proteine mit *M. musculus* TERRA- zu identifizieren. Um die angereicherten TERRA-interagierenden Proteine zu bestimmen, wurden identifizierte und quantifizierte Proteine mit denen verglichen, die bei der Verwendung einer Kontrollsequenz erhalten wurden. Die funktionelle Analyse der angereicherten Proteine zeigte eine Beteiligung von TERRA-interagierenden Proteinen am RNA-Stoffwechsel, der DNA-Replikation, der Mitose und der Chromatinorganisation. Weiterhin wurden die Daten mit öffentlich verfügbaren Datensätzen von *H. sapiens* TERRA-interagierenden Proteinen verglichen. Damit wurde gezeigt, dass trotz einer unterschiedlichen genomischen Herkunft die Funktionen von *M. musculus* und *H. sapiens* konserviert sind.

Zusammenfassend untersuchte dieses Projekt den unterschiedlichen Ursprung nichtkodierender RNA am Beispiel der TERRA-Moleküle in *M. musculus* und verdeutlichte die funktionale Konservierung von *H. sapiens* und *M. musculus* TERRA-interagierenden Proteinen.

3.1.3 Statement of contribution

This is a collaboration project where we conducted quantitative proteomics experiments. Marion Scheibe did all the proteomics data collection, including the RNA pull-downs and mass spectrometry sample processing, while I performed all the proteomics data analysis. This included the MS raw-data processing, the identification of enriched proteins, and the functional analysis. I also assembled and finalised the quantitative proteomics figure for the manuscript. Together with Marion Scheibe and Falk Butter, I also contributed to writing the quantitative proteomics section of the manuscript.

Supervision confirmation

Falk Butter

PAR-TERRA is the main contributor to telomeric repeat-containing RNA transcripts in normal and cancer mouse cells

NIKENZA VICECONTE,^{1,5} AXELLE LORIOT,^{1,5} PATRÍCIA LONA ABREU,² MARION SCHEIBE,³ ALBERT FRADERA SOLA,³ FALK BUTTER,³ CHARLES DE SMET,¹ CLAUD M. AZZALIN,² NAUSICA ARNOULT,⁴ and ANABELLE DECOTTIGNIES¹

¹Genetic and Epigenetic Alterations of Genomes, de Duve Institute, Université catholique de Louvain (UCLouvain), 1200 Brussels, Belgium

²Instituto de Medicina Molecular João Lobo Antunes, Faculdade de Medicina da Universidade de Lisboa, 1649-028 Lisboa, Portugal

³Quantitative Proteomics, Institute of Molecular Biology (IMB), 55128 Mainz, Germany

⁴MCBD-University of Colorado Boulder, Boulder, Colorado 80309-0347, USA

ABSTRACT

Telomeric repeat-containing RNA (TERRA) molecules play important roles at telomeres, from heterochromatin regulation to telomerase activity control. In human cells, TERRA is transcribed from subtelomeric promoters located on most chromosome ends and associates with telomeres. The origin of mouse TERRA molecules is, however, unclear, as transcription from the pseudoautosomal PAR locus was recently suggested to account for the vast majority of TERRA in embryonic stem cells (ESC). Here, we confirm the production of TERRA from both the chromosome 18q telomere and the PAR locus in mouse embryonic fibroblasts, ESC, and various mouse cancer and immortalized cell lines, and we identify two novel sources of TERRA on mouse chromosome 2 and X. Using various approaches, we show that PAR-TERRA molecules account for the majority of TERRA transcripts, displaying an increase of two to four orders of magnitude compared to the telomeric 18q transcript. Finally, we present a SILAC-based pull-down screen revealing a large overlap between TERRA-interacting proteins in human and mouse cells, including PRC2 complex subunits, chromatin remodeling factors, DNA replication proteins, Aurora kinases, shelterin complex subunits, Bloom helicase, Coilin, and paraspeckle proteins. Hence, despite originating from distinct genomic regions, mouse and human TERRA are likely to play similar functions in cells.

Keywords: telomeric RNA; PAR; TERRA interactome; telomere

INTRODUCTION

While telomeres have long been recognized as heterochromatic structures, they are active transcription sites. The first report of the existence of a transcriptional activity at telomeres dates back to 1989 with work from Rudenko and van der Ploeg (1989) in *Trypanosoma brucei*. A few years later, transcription was also observed at telomeres of bird lampbrush chromosomes (Solovei et al. 1994).

The first evidences of telomeric transcription in human and mouse cells were provided nearly 15 yr later (Azzalin et al. 2007; Schoeftner and Blasco 2008). Telomeric Repeat-containing RNA (TERRA) molecules were found to be transcribed from the C-rich telomeric DNA strand

and to be detectable with $(CCCTAA)_n$ probes in RNA fluorescence in situ hybridization (FISH). From the beginning, however, it emerged that mouse and human TERRA molecules may present some differences as their respective FISH patterns were quite distinct. In human interphase cells, TERRA-FISH signals were clearly detected at telomeres (Azzalin et al. 2007), while, in mouse cells, main TERRA-FISH signals appeared as only one or two large foci colocalizing with the q end of X/Y chromosomes (Schoeftner and Blasco 2008; Schoeftner et al. 2009; Zhang et al. 2009). On rare occasions (2–3 foci per nucleus), much smaller TERRA foci were also found to colocalize with mouse telomeres (López de Silanes et al. 2014).

⁵These authors contributed equally to this work.

Corresponding authors: anabelle.decottignies@uclouvain, nausica.arnoult@colorado.edu

Article is online at <http://www.majournal.org/cgi/doi/10.1261/ma.076281.120>.

© 2021 Viceconte et al. This article is distributed exclusively by the RNA Society for the first 12 months after the full-issue publication date (see <http://majournal.cshlp.org/site/misc/terms.xhtml>). After 12 months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

In line with human TERRA detection at telomeres by RNA-FISH, many studies clearly demonstrated that human chromosome ends are transcribed from subtelomeric promoters that are located directly upstream of telomeric repeats (Nergadze et al. 2009; Deng et al. 2012; Porro et al. 2014a; Diman et al. 2016; Feretzaki and Lingner 2017; Koskas et al. 2017; Sagie et al. 2017; Feretzaki et al. 2019). A possible explanation to account for the distinct TERRA-FISH patterns between human and mouse cells was recently provided by RNA-seq experiments after TERRA capture from mouse ESC (Chu et al. 2017a). That study revealed the existence of a novel TERRA species, dubbed PAR-TERRA, originating from the pseudoautosomal locus of Xq/Yq subtelomeric regions, and accounting for more than 99% of all TERRA transcripts in mouse ESC. A few reads emanating from a small subset of telomeres were also identified in that study and included the previously reported Telo 18q TERRA (López de Silanes et al. 2014). The transcript emanating from Telo 18q was, however, about 4000-fold less abundant than PAR-TERRA in ESC (Chu et al. 2017a). Using the CHIRT technique, which combines ChIRP (Chromatin isolation by RNA purification) and CHART (Capture hybridization analysis of RNA targets), the same study also revealed that PAR-TERRA binds *in trans* to most chromosome ends and is the major (UUAGGG)_n-containing RNA species associating with mouse ESC telomeres. Altogether, these observations questioned the telomeric origin of mouse TERRA, as recently discussed (Diman and Decottignies 2018).

Here, we used several approaches, including the interrogation of publicly available RNA-seq data sets, to further investigate the genomic origin of mouse TERRA in mouse embryonic fibroblasts (MEF), ESC, normal tissues, and cancer cells. Our results clearly indicated that, similarly to the previous observations in ESC, PAR-TERRA molecules account for the vast majority of TERRA transcripts in all mouse cell types. We also identified a new TERRA species transcribed from chromosome 2. However, both this newly identified Chr 2 TERRA and the Telo 18q TERRA appear to minimally contribute to the total cellular pool of (UUAGGG)_n repeats. To further probe the conservation of mouse and human TERRA, we performed a TERRA interactome screen in mouse cells us-

ing a previously published SILAC-based pull-down approach (Scheibe et al. 2013). Cross-comparison of our results with an *in vivo* proteomic screen for TERRA-interacting proteins (Chu et al. 2017b) revealed that, despite distinct genomic origins and distinct nucleotide sequences, human and mouse TERRA are likely to interact with similar cellular proteins.

RESULTS

TERRA-FISH foci colocalize with telomeres and their intensity is proportional to telomere length in human cells

In human primary and cancer cells, most TERRA-FISH foci colocalize with the telomeres of interphase nuclei (Fig. 1A; Azzalin et al. 2007; Arnoult et al 2012; Diman et al. 2016). Consistent with previous northern blot experiments showing that the length of human TERRA molecules is proportional to telomere length (Yehezkel et al. 2008; Arnoult

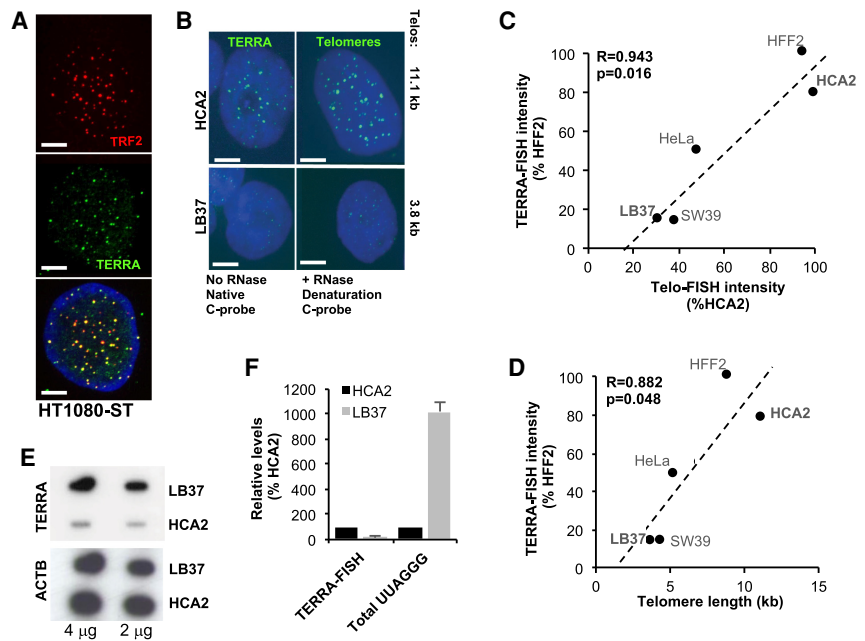


FIGURE 1. TERRA-FISH signals correlate with telomere length but not with UUAGGG content in human cells. (A) Immunofluorescence against TRF2 (red) combined with RNA-FISH to detect TERRA with a (CCCTAA)₇ LNA green probe in HT1080-ST human cell line. DNA is stained with DAPI (blue). Scale bar, 5 μm. (B) Comparison between TERRA-FISH (left) and telomeric FISH (right) in HCA2 (telomere length: 11.1 kb) and LB37 (telomere length: 3.8 kb) human cells using the (CCCTAA)₇ LNA green probe and the same exposure time. Experimental conditions are indicated below. Scale bar, 5 μm. (C) Correlation between total nuclear TERRA-FISH intensity (expressed as % of intensity in HFF2 cells) and total Telo-FISH intensity (expressed as % of intensity in HCA2 cells). At least 50 nuclei were quantified for TERRA-FISH or Telo-FISH in each cell line. Cell line names are indicated. R = 0.943; P two-tailed = 0.016. (D) Correlation between TERRA-FISH intensity (expressed as % of intensity in HFF2 cells) and telomere length (kb) evaluated by TRF (Tilman et al. 2009; Arnoult et al. 2012). R = 0.882; P two-tailed = 0.048. (E) RNA slot-blots in LB37 and HCA2 human cells. Total RNA (2 or 4 μg) isolated from LB37 or HCA2 cells was hybridized with either the (CCCTAA)₄ probe or a probe against human ACTB. (F) Quantification from B and E. Data are normalized to HCA2. Mean ± S.D.

et al. 2012; Van Beneden et al. 2013), we confirmed, by TERRA-FISH, that fluorescence signal intensity mirrors telomeric DNA FISH intensity and, therefore, telomere length (Fig. 1B–D). The intensity of TERRA-FISH foci, however, does not reflect the total levels of *UUAGGG* repeats in human cells, as shown by slot-blot hybridization of total RNA isolated from HCA2 normal fibroblasts and LB37 lung cancer cells which, because of subtelomeric promoter hypomethylation (Nergadze et al. 2009; Diman et al. 2016; Sagie et al. 2017; Feretzaki et al. 2019; Le Berre et al. 2019), strongly up-regulate telomeric transcription (Fig. 1E,F). Taken together, these observations suggest that a substantial fraction of the human TERRA pool escapes detection by RNA-FISH and that the telomere-bound molecules may likely be the only ones that are detected by RNA-FISH. The reasons for this are still unclear but may derive from the preextraction step of RNA-FISH protocol, which would wash out soluble TERRA and leave chromatin-bound TERRA as the only detectable species. It is also possible that soluble TERRA is not efficiently recognized by the RNA-FISH probe because it is folded into tight secondary structures such as G-quadruplexes (G4), or heavily bound to RNA-binding proteins.

In mouse cells, most TERRA-FISH signals do not colocalize with telomeres and their intensity does not correlate with telomere length

We next applied RNA-FISH to mouse cell lines with various telomere lengths in order to test whether signal intensity may similarly be proportional to telomere length. To do this, we first evaluated telomere length by TRF in the following cell lines: 3T3 (spontaneously immortalized MEF), L929 (spontaneously immortalized adult and adipose tissue-derived fibroblasts), RAW264.7 (Abelson murine leukemia virus transformed macrophages), J774A.1 (reticulum cell sarcoma), M1 (myeloblast cell line), and Neuro-2a (neuroblastoma) (Fig. 2A). We selected four mouse cell lines with either short (J774A.1, 4.5 kb), average (M1, 5.8 kb) or long (L929, 9.5 kb and Neuro-2a, 19.9 kb) telomeres for TERRA-FISH experiments and performed RNase A treatments as controls (Fig. 2B). Our data showed that the total TERRA-FISH signal intensity does not correlate with telomere length in mouse cells ($P=0.53$) (Fig. 2C). As reported earlier, the RNA-FISH profiles are also very different from the ones obtained in human cells, with the presence of 2–3 prominent TERRA foci per nucleus (Fig. 2B,D). When TERRA-FISH signals were intense, like in J774A.1 and Neuro-2a cell lines, additional, but much smaller foci were detected using the same exposure conditions (Fig. 2B). To further investigate the colocalization of TERRA foci with mouse telomeres, we combined RNA-FISH with immunofluorescence against the shelterin protein Terf1 in the Neuro-2a cell line. In sharp contrast with the consistent colocalization of TERRA-FISH signals with the other

shelterin protein TRF2 in human cells (Fig. 1A), and in line with previous reports (López de Silanes et al. 2014; Chu et al. 2017b), only 2–4 small TERRA-FISH foci overlapped with Terf1 in mouse interphase nuclei (Fig. 2D). Conversely, in the same Neuro-2a cell line, Terf1 signals were colocalizing with telomeres in control DNA-FISH experiments using the same probe (Fig. 2E). The lack of correlation between telomeric DNA and TERRA-FISH signal intensity, together with the low frequency of colocalization events between TERRA and telomeres, suggest that telomere transcription may not be the main source of TERRA in mouse cells.

Similarly to human cells, however, we found that *UUAGGG* levels, normalized to *ACTB* mRNA (Fig. 2F), do not correlate with *TTAGGG* levels in mouse cells (Fig. 2G; Supplemental Fig. S1A,B, $P=0.28$). This could be explained either by a nontelomeric origin of mouse TERRA or, like in human cells, by a distinct transcriptional activity at telomeres of various mouse cell lines.

Altogether, our experiments suggest very distinct regulations of TERRA in human and mouse cells (Fig. 2H), reinforcing the idea that the origin and molecular features of TERRA molecules may be very distinct in these two species.

TERRA molecules are produced from various mouse genomic loci

Based on the above-described findings, we hypothesized that $(UUAGGG)_n$ -containing sequences may result from the transcription of intrachromosomal *TTAGGG*-rich sequences. To identify these putative loci, we first scanned successive 2 kb-long regions of the mouse genome for the presence of at least 30 telomeric repeats—not necessarily consecutive—including either pure (*TTAGGG*) or degenerate (*TAAGGG*, *TGAGGG*, *TTGGGG*, *GTAGGG*, or *TCAGGG*) motifs (Fig. 3A). The arbitrary cut-off of 30 telomeric repeats was based on our preliminary analysis of the mouse genome, which suggested that the criteria was stringent enough to detect telomeric repeat-enriched regions above the background. A total of 105 intrachromosomal loci containing ≥ 30 telomeric motifs were identified. In a second step, six RNA-seq data sets with high sequencing coverage (>140 million paired reads) from ES, forebrain, frontal lobe, B-cell lymphoma, MEF, and MEL leukemia cells were selected for alignment with the identified intrachromosomal telomeric repeat-containing genomic loci. Using the same RNA-seq data sets, and whenever subtelomeric sequences were available, telomeric transcripts were also searched for by screening for subtelomeric reads directly adjacent to telomeres. As previously reported (López de Silanes et al. 2014), we detected few reads from the 18q subtelomere (Fig. 3B,C). A very limited number of reads appeared to similarly emanate from the 10q subtelomere, but not from any other

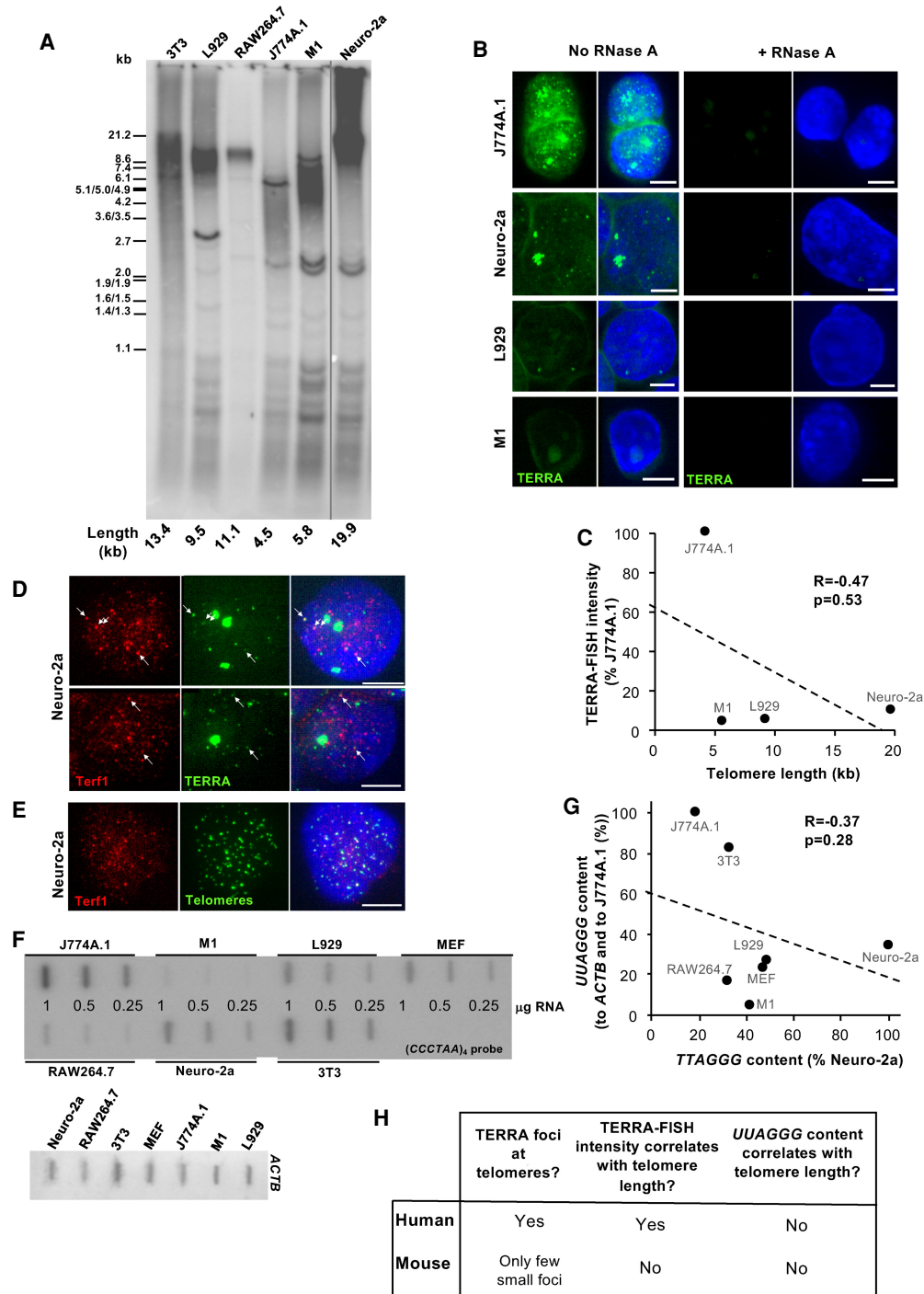


FIGURE 2. TERRA-FISH signals correlate with UUAGGG content but not with telomere length in mouse cells. (A) TRF analysis on genomic DNA isolated from the indicated mouse cell lines. Position of the ladder is indicated on the left with the corresponding size (kb). Telomere length, calculated with Telotool software, is indicated below. (B) TERRA-FISH in J774A.1, Neuro-2a, L929, and M1 mouse cell lines using the (CCCTAA)₇ LNA green probe with (right) or without (left) RNase A treatment. DNA is stained with DAPI (blue). Bar scale, 5 μm. (C) Correlation between measurements in A and B. At least 35 nuclei were quantified for TERRA-FISH in each cell line. $R = -0.47$; P two-tailed = 0.53. (D) Immunofluorescence against Terf1 (red) combined with RNA-FISH to detect TERRA with a (CCCTAA)₇ LNA green probe in Neuro-2a mouse cells. Arrows indicate colocalization events. DNA is stained with DAPI (blue). Bar scale, 5 μm. (E) Immunofluorescence against Terf1 (red) combined with FISH to detect telomeres with a (CCCTAA)₇ LNA green probe in Neuro-2a mouse cells. DNA is stained with DAPI (blue). Bar scale, 5 μm. (F) RNA slot-blots in the indicated mouse cell lines. Total RNA (1, 0.5, or 0.25 μg) was hybridized with the (CCCTAA)₄ probe (upper panel) or with a probe against mouse ACTB (lower panel, 0.5 μg). (G) Correlation between measurements from F and TTAGGG content evaluated by slot-blot (Supplemental Fig. S1A, B). $R = -0.37$; P two-tailed = 0.28. (H) Summary of the differences between human and mouse TERRA.

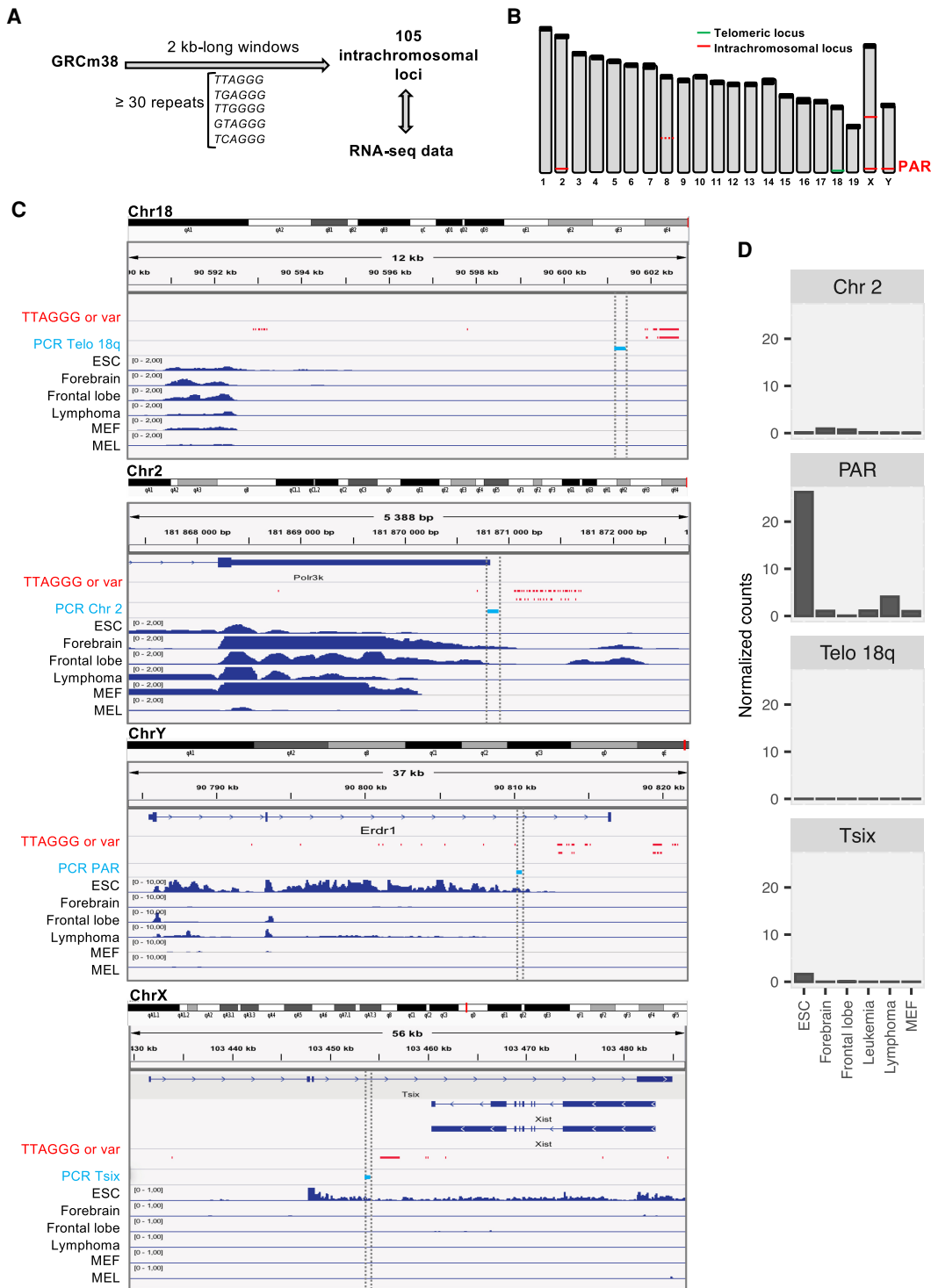


FIGURE 3. In silico search for candidate mouse genomic loci producing UUAGGG-containing RNAs. (A) Workflow for the identification of mouse candidate genomic loci for TERRA production. (B) Graphical representation of mouse chromosomes and putative loci for TERRA production in green (telomeric loci) or red (intrachromosomal loci). (C) IGV screenshots showing coverage density over regions of interest for the indicated mouse cell lines or tissues. The red line on each chromosome indicates the region displayed *below*. Genes or portions of genes are represented, with dark blue boxes corresponding to exons and dark blue lines with arrows corresponding to introns. TTAGGG or variant (as indicated in A) repeats are indicated in red. Regions corresponding to PCR products in the study appear in light blue with dotted lines on both sides. Displayed coverages are normalized (each value was multiplied by 1×10^6 and divided by the total number of reads) and correspond to reads unambiguously mapped and resulting from transcription in sense orientation exclusively. (D) Quantification of reads aligned to the regions shown in C (“PCR products”). Counts are normalized (read counts were divided by the region length [kb] and multiplied by 1×10^6 /total number of reads).

chromosome ends with available sequence. Analysis of reads produced from the 105 identified telomeric repeat-containing intrachromosomal loci revealed the possible contribution of two loci to the production of $(UUAGGG)_n$ -containing RNAs in all cell/tissue types: one on Chr 2 (3'UTR of *Polr3k* gene) and one on Chr X/Y (PAR locus-*Erdr1* gene) (Fig. 3B,C; Supplemental Fig. S2). The *Tsix* locus on Chr X was identified as an additional intrachromosomal source of $(UUAGGG)_n$ -containing transcripts in mouse ES cells exclusively (Fig. 3B,C; Supplemental Fig. S2) and is known to be exclusively transcribed from the future active X chromosome in female cells. Although not detected in ES cells, a fourth intrachromosomal locus on Chr 8 may (Fig. 3B, red dotted line) also contribute to TERRA production in some cells, including brain cells. This region of Chr 8 was however not considered as a strong contributor to overall TERRA production in mouse tissues as the total number of *TTAGGG* repeats was only 31 and the expression of the host gene, *Inpp4b*, is restricted to some tissue types only (not shown). The other candidate loci showed no or extremely low count number and were not selected as good candidates for contributing to TERRA production.

The pseudoautosomal PAR locus was recently identified as the main source of TERRA in mouse ES cells and, through CHIRT analysis, PAR-TERRA was found to bind *in trans* to various genomic loci (Chu et al. 2017a). Careful analysis of the CHIRT data revealed that, while most PAR-TERRA and TERRA peaks indeed overlap in the mouse genome, PAR-TERRA is not detected at *Polr3k* 3'UTR locus, where TERRA peaks are readily detected (Chu et al. 2017a), suggesting that a TERRA molecule, unrelated to PAR-TERRA, is produced from and binds to this locus. Together with our analysis, this shows that Chr 2 TERRA is a newly identified source of mouse TERRA.

To compare the relative expression levels of the four major TERRA transcripts (Telo 18q, Chr 2, PAR, and *Tsix*) in the selected RNA-seq data sets, we quantified the reads emanating from small unique regions located upstream of the $(TTAGGG)_n$ repeats (Fig. 3C, light blue boxes). Strand specificity was taken into account to exclusively quantify the reads corresponding to $(UUAGGG)_n$ -containing transcripts. The analysis revealed that PAR-TERRA was the most abundant species in ESC, MEF, B-cell lymphoma and MEL leukemia cell line, while similar levels of Chr 2 TERRA and PAR-TERRA were measured in brain tissues (Fig. 3D). Importantly, Telo 18q TERRA was barely detectable in all samples, suggesting that the contribution of this locus to mouse TERRA is extremely low (Fig. 3D).

Candidate mouse TERRA molecules contain *UUAGGG* repeats

To validate the production of $(UUAGGG)_n$ -containing sequences from the four genomic candidate loci, we designed primers located upstream of the *TTAGGG*

repeats of each locus. For Telo 18q, primers are located within the possible 3'UTR of *LOC108168395* downstream from *Tmx3* gene, within the last 20 kb before 18q telomeric repeats (Fig. 4A). For Chr 2, primers are in the 3'UTR region of the *Polr3k* gene. Primers for PAR and *Tsix* loci are located within introns (Fig. 4A). Primer efficiency was tested (Supplemental Fig. S3) and two distinct reverse transcriptions (RT) were performed: one with random hexamers and one with $(CCCTAA)_5$ primers (Telo primers). If the candidate transcript contains long stretches of *UUAGGG* repeats, we expect a strong enrichment of the corresponding cDNA when RT is performed with Telo primers. If, on the other hand, the transcript contains fewer and/or interspaced *UUAGGG* repeats, the enrichment should be weaker. Finally, if the transcript is completely devoid of *UUAGGG* repeats, we expect the efficiency of the RT with Telo primers to be strongly reduced and the corresponding cDNA to be virtually absent.

Total RNA was isolated from J774A.1 and ES cells. Consistent with the respective abundance of consecutive *TTAGGG* repeats at the tested loci (Supplemental Fig. S2), we observed either a strong (Telo 18q, Chr 2) or a moderate (PAR, *Tsix*) enrichment of cDNA species when Telo primers were used for the RT instead of random primers (Fig. 4B). Importantly, the levels of *Sod1*, used as negative control for $(UUAGGG)_n$ -containing mRNA, were at background levels when Telo primers were used for RT (Fig. 4B). Altogether, these data confirm the presence of *UUAGGG* repeats in Telo 18q, Chr 2, PAR, and *Tsix* TERRA species.

PAR-TERRA is the main source of *UUAGGG*-rich RNA molecules in ESC, MEF, and immortalized mouse cell lines

As the candidate TERRA loci likely produce RNA molecules with distinct *UUAGGG* repeat contents, we next sought to determine which of those TERRA loci is the main contributor to the pool of *UUAGGG* repeats in mouse cells. To do this, we performed RT-qPCR analyses on random primer-synthesized cDNA in MEF, ES cells and in the six mouse cell lines, and compared these measurements to *UUAGGG* levels measured by slot-blots. Our RT-qPCR analyses confirmed that PAR-TERRA levels are much higher than the other tested transcripts in all cell lines (Fig. 5A). The strong correlation ($R = 0.967$, $P = 0.0001$) that we obtained between PAR-TERRA levels and total *UUAGGG* levels, measured by slot-blot, further confirmed that PAR-TERRA is the main source of *UUAGGG* repeats in mouse cells (Fig. 5B,C). Conversely, the levels of either Telo 18q ($R = -0.49$, $P = 0.223$) or Chr 2 ($R = -0.25$, $P = 0.554$) did not correlate with total *UUAGGG* levels (Fig. 5C).

A short stretch of *TTAGGG* repeats located between the *LOC108168395* locus and the 18q telomere (Supplemental Fig. S4A) was previously proposed as a possible source

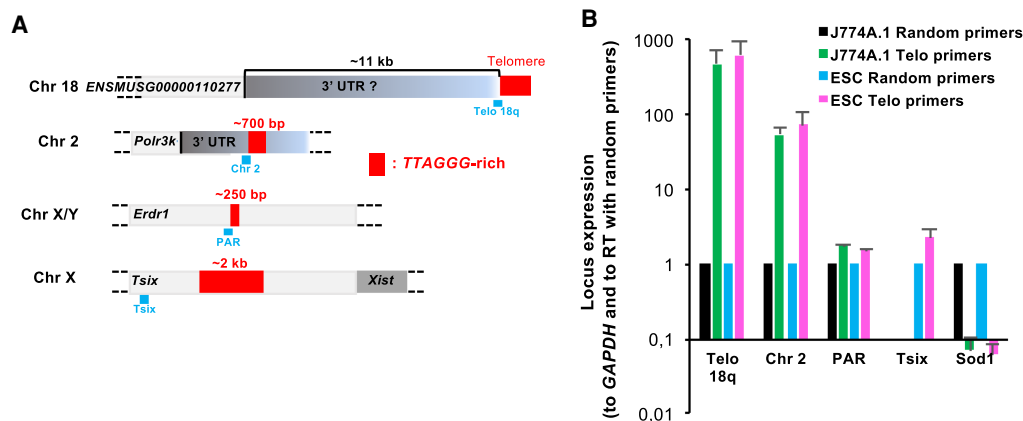


FIGURE 4. Candidate TERRA molecules contain *UUAGGG* repeats. (A) Graphical representation of the genomic context for the four mouse TERRA loci analyzed in this study. *TTAGGG*-rich regions, whether at telomeres or at intrachromosomal loci, are shown by red boxes and their length is indicated. PCR product positions are indicated with light blue boxes. (B) Relative expression levels of Telo 18q, Chr 2, PAR, or Tsix TERRA loci and *Sod1* in cDNA from J774A.1 and ES cells synthesized with either random or Telo and GAPDH primers. Values are normalized first to GAPDH and then to the ratio measured in J774A.1 cells when random primers are used for RT. Mean \pm S.D.

for TERRA (López de Silanes et al. 2014). However, our data failed to reveal any correlation between this subtelomeric 18q transcript and the relative *UUAGGG* levels in the tested cell lines (Supplemental Fig. S4B).

Telomere deprotection through Terf2 depletion does not up-regulate telomere transcription

Previous work in human cells revealed that TERRA transcription is up-regulated from TRF2-depleted telomeres (Caslini et al. 2009; Porro et al. 2014a,b). To test the hypothesis that telomere deprotection may be similarly associated with enhanced transcription of mouse telomeres, we extracted RNA from SV40-immortalized MEF conditionally knocked-out for Terf2 (Fig. 6A). We also designed new pairs of primers for 5q and 11q chromosome ends located at 15 bp and 10 bp from the telomeric tract, respectively (Fig. 6B). Primers specifically amplified genomic DNA from J774A.1 cells (Fig. 6B; Supplemental Fig. S3); however, no amplification product was obtained with J774A.1 cDNA as template (Fig. 6B). Similarly, no product was obtained for 5q or 11q after 40 cycles of PCR on cDNA from Terf2 F/+ or tamoxifen-treated Terf2 F/F MEF, although Telo primers were used for the RT (Fig. 6C). In addition, 18q TERRA levels were not up-regulated upon Terf2 depletion (Fig. 6C), suggesting, once again, a very distinct regulation of telomere transcription in human and mouse cells.

Comparison of human and mouse TERRA interactomes

Despite the distinct chromosomal origin of mouse and human TERRA molecules, important cellular functions, in-

cluding heterochromatin regulation (Bettin et al. 2019), appear to be exerted by TERRA in the two organisms. To better evaluate the extent of conservation of TERRA-associated functions in mouse and human, we performed a SILAC-based in vitro purification with a $(UUAGGG)_8$ biotinylated TERRA-like oligonucleotide as a bait to identify mouse TERRA-interacting proteins (Scheibe et al. 2013). Even though PAR-TERRA appears to be the most abundant TERRA species in mouse cells, and does not consist of long uninterrupted $(UUAGGG)_n$ sequences, we believe that the use of a $(UUAGGG)_8$ biotinylated probe is appropriate, as a number of PAR-TERRA/protein interactions likely involve *UUAGGG* repeats (Chu et al. 2017b).

We incubated mouse R1/E ES cell extracts with a $(UUAGGG)_8$ biotinylated TERRA probe or with a biotinylated $(GUGUGA)_8$ probe as a control for specificity. A total of 307 candidate proteins were identified that showed at least a fourfold enrichment over the control probe (Supplemental Fig. S5A,B) and the number increased to 581 with a threshold of twofold (Supplemental Table S1). To help identify functional classes of mouse TERRA-interacting proteins, we compared our data set of 307 proteins with the one previously obtained through iDRiP (identification of Direct RNA interacting Proteins) in mouse ESC (134 proteins) (Chu et al. 2017b). Thirty candidates were common to both data sets (yellow rectangles in Fig. 7A) and additional 98 proteins, from either screen (candidates from the SILAC screen are shown in bold), belonged to similar protein families, leading to a total of 128 “common” candidate proteins (Fig. 7A). A STRING analysis was performed on the 128 proteins and revealed possible roles for mouse TERRA in chromatin remodeling, RNA metabolism, DNA replication, ribosome biogenesis, or mitosis, with an enrichment in proteins related to centromeres,

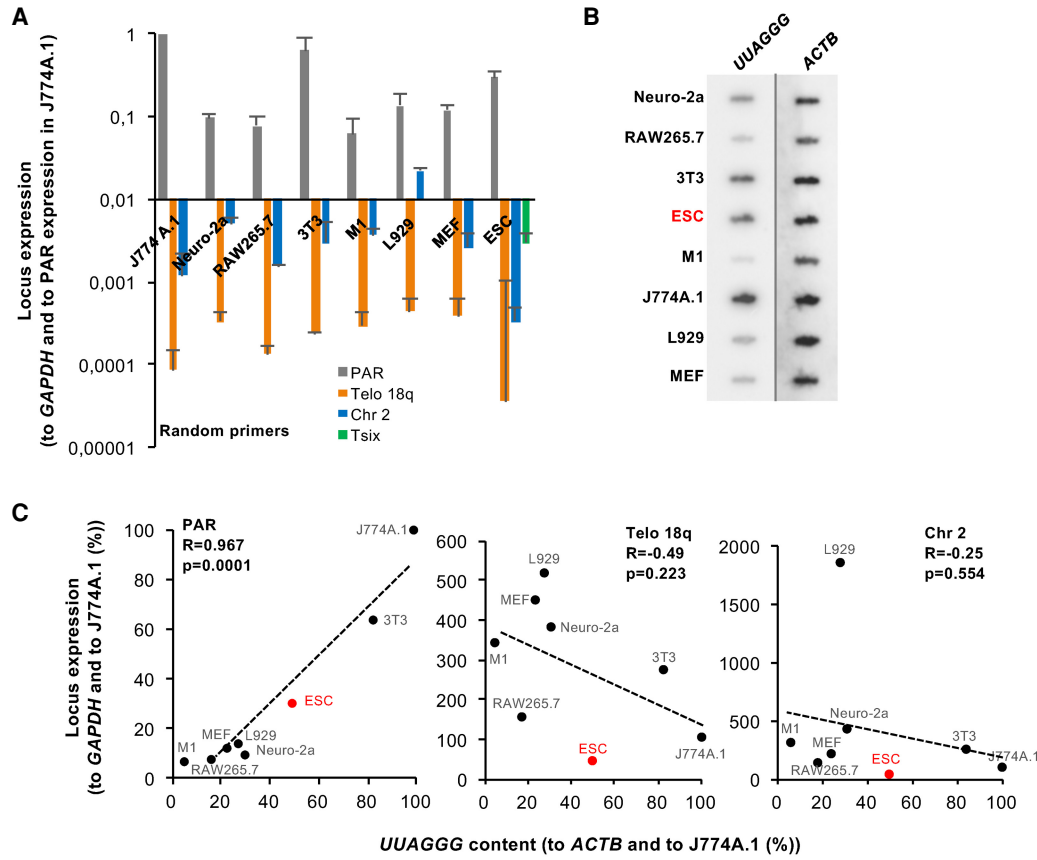


FIGURE 5. TERRA transcripts mostly originate from the PAR locus in mouse cells. (A) Relative expression levels of PAR, Telo 18q, Chr 2, or Tsix TERRA loci in cDNA from the indicated cell lines synthesized with random primers. Values are normalized first to GAPDH and then to the PAR expression level in J774A.1 cells. Mean \pm S.D. (B) RNA slot-blot in the indicated mouse cell lines. Total RNA (1 μ g) was hybridized with the (CCCTAA)₄ probe or with a probe against mouse ACTB. (C) Correlations between measurements from A and B as indicated (all measurements were normalized to J774A.1). R values and P two-tailed values are indicated for each graph.

chromatid cohesion, mitotic spindle or cytokinesis (Fig. 7A). Consistent with the previously reported interaction between mouse or human TERRA and the PRC2 complex (Bettin et al. 2019), our pull-down approach recovered the mouse Eed, Suz12, Jarid2 and Mtf2 subunits, with Mtf2 being shared with the iDRiP screen (Fig. 7A). The Terf1 subunit of the shelterin complex was also recovered in both screens. Similarly, we confirmed the interaction of TERRA with many mouse proteins involved in RNA metabolism, including proteins from the hnRNP (heterogeneous nuclear ribonucleoprotein) family, as initially demonstrated by López de Silanes et al. (2010). Other interacting proteins included Orc1, Orc2 and Blm—with roles in DNA replication and DNA damage repair—the Cajal body protein Coilin and a series of proteins with various functions in mitosis, like Aurora kinases b and c, the inner centromere protein Incenp, Sgol2, involved in sister chromatid cohesion regulation, Tpx2, with a key role in spindle assembly during mitosis and Mki67, that associates with mitotic chromosomes (Fig. 7A). As the in vivo iDRiP screen mostly monitored the interactions between PAR-TERRA and

mouse ES cell proteins, the consistency that we observed with the SILAC pull-down screen agrees with interactions happening through the UUAGGG repeats. To further validate our SILAC pull-down, we performed RNA immunoprecipitation (RIP) experiments using extracts from immortalized MEF and antibodies against two identified candidates, Suz12 and Blm. Because TERRA interacts with human TRF2 both in vitro and in cells (Deng et al. 2009; Lee et al. 2018), we also performed RIP using anti-Trf2 antibodies even though the protein was not recovered in our SILAC-pull down experiments. In line with previous RIP experiments in mouse iPS cells (Marión et al. 2019), PAR-TERRA coimmunoprecipitated with the anti-Suz12 antibody (Fig. 7B). Further validating our pull-downs, PAR-TERRA was also recovered using antibodies against Blm (Fig. 7B). Finally, PAR-TERRA was also detected in fractions isolated using the anti-Trf2 antibodies. In all pull-downs, Telo 18q RNA was barely detectable (Fig. 7C). Together, these results confirm the validity of our method and clearly establish that PAR-TERRA and human TERRA bind to similar sets of proteins. The differences in

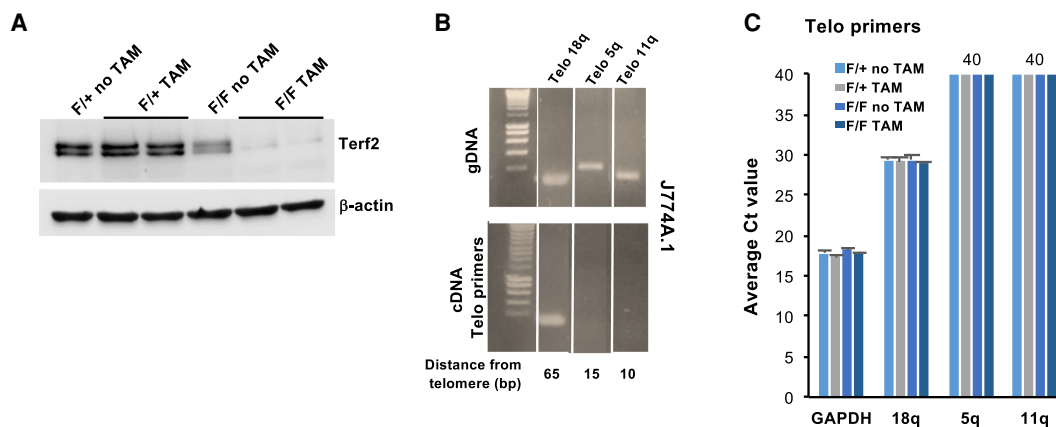


FIGURE 6. Transcription of mouse telomeres is not increased upon Terf2 depletion. (A) Western blot analysis of Terf2 in total cell extracts from CreERT2 Terf2 F/+ and Terf2 F/F SV40T-immortalized MEF. β -actin is shown as loading control. (B) Visualization of PCR products from Telo 18q, Telo 5q, and Telo 11q obtained through amplification on either gDNA (upper panel) or cDNA (lower panel) prepared from J774A.1 cells. The distance between each PCR product and the respective telomere is indicated below (bp). (C) C_t values from RT-qPCR experiments on cDNA synthesized with Telo and GAPDH primers from the F/+ or F/F fibroblast cell lines treated as indicated. Mean \pm S.D.

the amounts of PAR-TERRA measured in the three different RIP do not necessarily reflect the strength of PAR-TERRA interaction with Suz12, Blm, and Terf2 due to different IP efficiencies. The fact that Terf2 was not detected in the SILAC screen might indicate that PAR-TERRA/Terf2 interactions are transient or not strong enough to be detected without cross-linking.

We next searched for common TERRA-interacting proteins in human and mouse cells by comparing our two SILAC-based pull-down screens, using a threshold of greater than or equal to twofold enrichment for both screens (Supplemental Tables S1, S2). Our analyses revealed that 188 out of the 454 (41%) human TERRA-interacting protein candidates are also present in the mouse SILAC screen while 188 out of the 581 mouse candidates (32%) are detected in the human screen (Fig. 8A; Supplemental Fig. S6A,B). The common interacting proteins are involved in RNA metabolism, DNA replication, mitosis (Aurora kinases a/b/c, Mcm5, Rfc3/4,.), or chromatin organization (including PRC2 complex) (Fig. 8B) and also include components of the CDC5L complex, the Cajal bodies or the paraspeckle subnuclear bodies.

Based on the large overlap of interacting proteins and associated pathways, it is conceivable that many TERRA-associated functions are shared between human and mouse cells.

DISCUSSION

From the first reports of TERRA in human and mouse cells, differences between the two species rapidly emerged as very distinct TERRA-FISH patterns were observed. While human TERRA was clearly detected at telomeres, colocalization of TERRA with mouse telomeres was rarely ob-

served by RNA-FISH in interphase cells (López de Silanes et al. 2014). More recently, CHIRT experiments revealed that a TERRA-like transcript, produced from the pseudoautosomal PAR locus of mouse ES cells, was able to bind, *in trans*, to most mouse telomeres (Chu et al. 2017a). Here, we confirmed the distinct origin of TERRA in human and mouse cells. Using the same probes and tools, and together with our previous studies (Azzalin et al. 2007; Arnoult et al. 2012; Van Beneden et al. 2013), we showed that TERRA results from the transcription of telomeres in human cells, while the vast majority of the *UUAGGG* repeats comes from the PAR locus in all the mouse cell lines that we tested. PAR-TERRA corresponds to an intronic region of the *Erd1* gene on the X/Y chromosome, and thus reinforces the idea that spliced introns can be precursors of noncoding RNAs, including microRNAs or long noncoding RNAs (Hesselberth 2013). Our RNA-FISH experiments using a telomeric probe in mouse cells revealed a good correlation between TERRA-FISH intensity and the level of PAR-TERRA molecules measured by RT-qPCR. This contrasts with human cells in which TERRA-FISH intensity does not correlate with the total TERRA levels as FISH signals are only visible at telomeres. We propose that telomere-bound human TERRA molecules can be detected by RNA-FISH while the other TERRA molecules escape detection, possibly because of G4 structures forming along the long (*UUAGGG*)_n tracts as previously shown (Biffi et al. 2012; Hirashima and Seimiya 2015) or because human TERRA, when not bound to telomeres, is lost during the preextraction step of the RNA-FISH protocol. PAR-TERRA, on the other hand, in light of its very distinct nucleotide sequence with interspaced telomeric repeats (Supplemental Fig. S2), may not fold into G4 structures and/or may interact more strongly with chromatin. In this

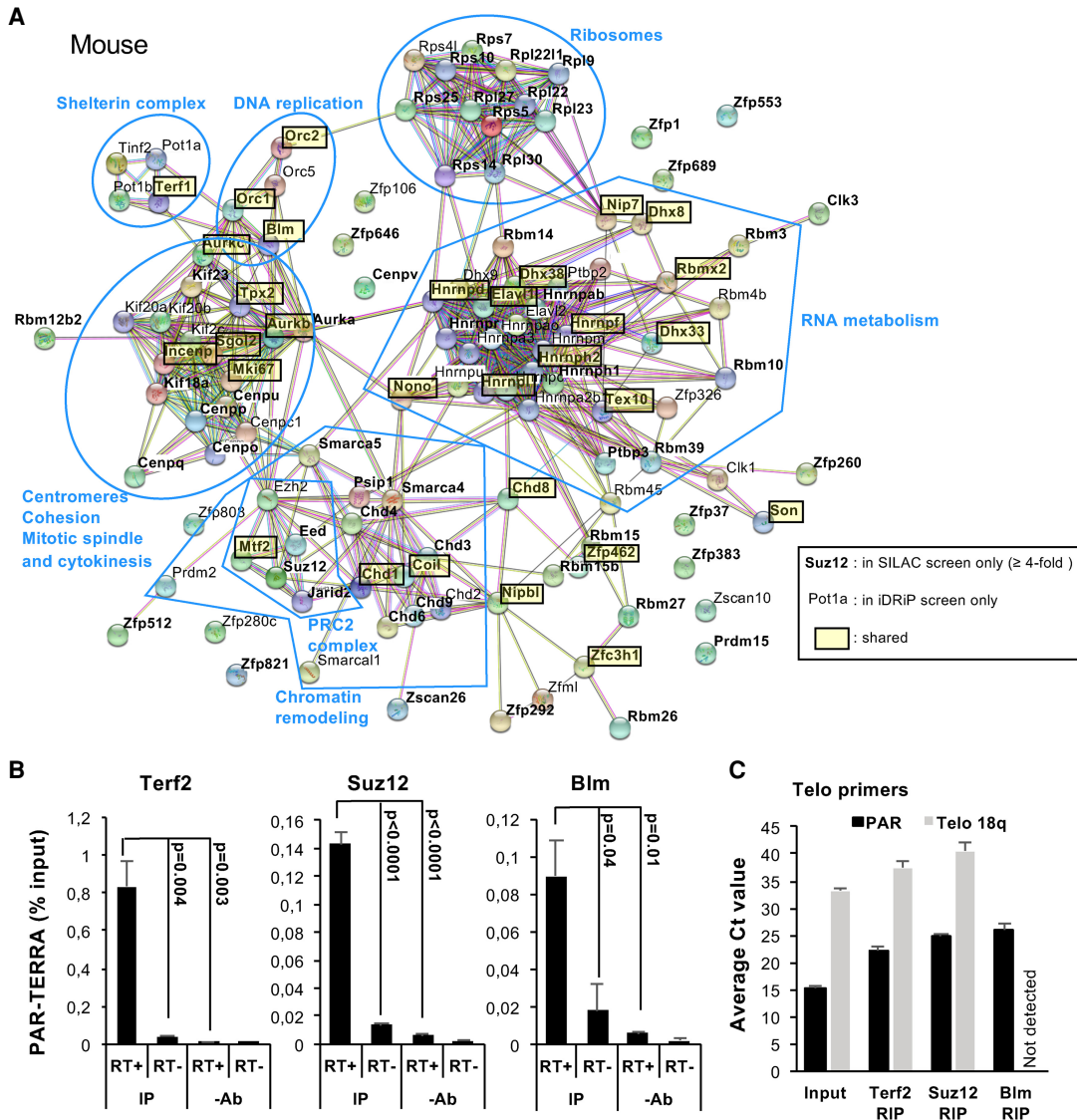


FIGURE 7. SILAC screening of the quantitative mouse TERRA pull-down. (A) STRING network showing the 128 mouse TERRA-interacting candidates from either the pull-down screen (using greater than or equal to fourfold enrichment as threshold) or the iDRiP screen (Chu et al. 2017b) that either showed perfect overlap between the two screens (highlighted in yellow) or were belonging to similar protein families. Candidates that were exclusively found in the SILAC pull-down screen are shown in bold. Line color of the edges indicates the type of interaction evidence between the nodes according to STRING default parameters. (B) Validation of Suz12 and Blm as PAR-TERRA-interacting proteins, using the shelterin complex as positive control for binding. Three independent RNA-IP experiments were performed for each protein using nuclear extracts from 2575i MEF and Telo primers for RT. No antibody was added in the control experiments (-Ab) and additional controls lacking reverse transcriptase (-RT) were performed to monitor gDNA contamination. PAR-TERRA transcript levels in the IP were normalized to input (%). Mean \pm S.E.M, two-tailed unpaired *t*-tests. (C) Average *C_t* values for PAR-TERRA and Telo 18q RT-qPCR in the input and the RIP samples. Mean \pm S.D., *n* = 3 independent experiments.

respect, CHIRT experiments in mouse ES cells clearly revealed the presence of PAR-TERRA binding sites throughout the genome and at the PAR locus (Chu et al. 2017a,b).

The in silico search for intrachromosomal mouse candidate TERRA loci unveiled a new locus on chromosome 2, likely located within the 3'UTR region of the *Polr3k* gene. Long noncoding RNA functions have previously been proposed for 3'UTR regions of genes, whether as whole molecules or as cleaved fragments (Mayr 2017). Based on our

analyses of RNA-seq data sets, this Chr 2 locus appears to contribute significantly to the pool of TERRA in some mouse tissues, including the forebrain and the frontal lobe. Interestingly, a careful analysis of the CHIRT data from Chu et al. (2017a) revealed the presence of a TERRA peak, but not a PAR-TERRA peak, at the *Polr3k* locus, that likely corresponds to the binding of Chr 2 TERRA molecules. This, in turn, suggests that Chr 2 TERRA binds to the locus where it is produced. Future work is needed

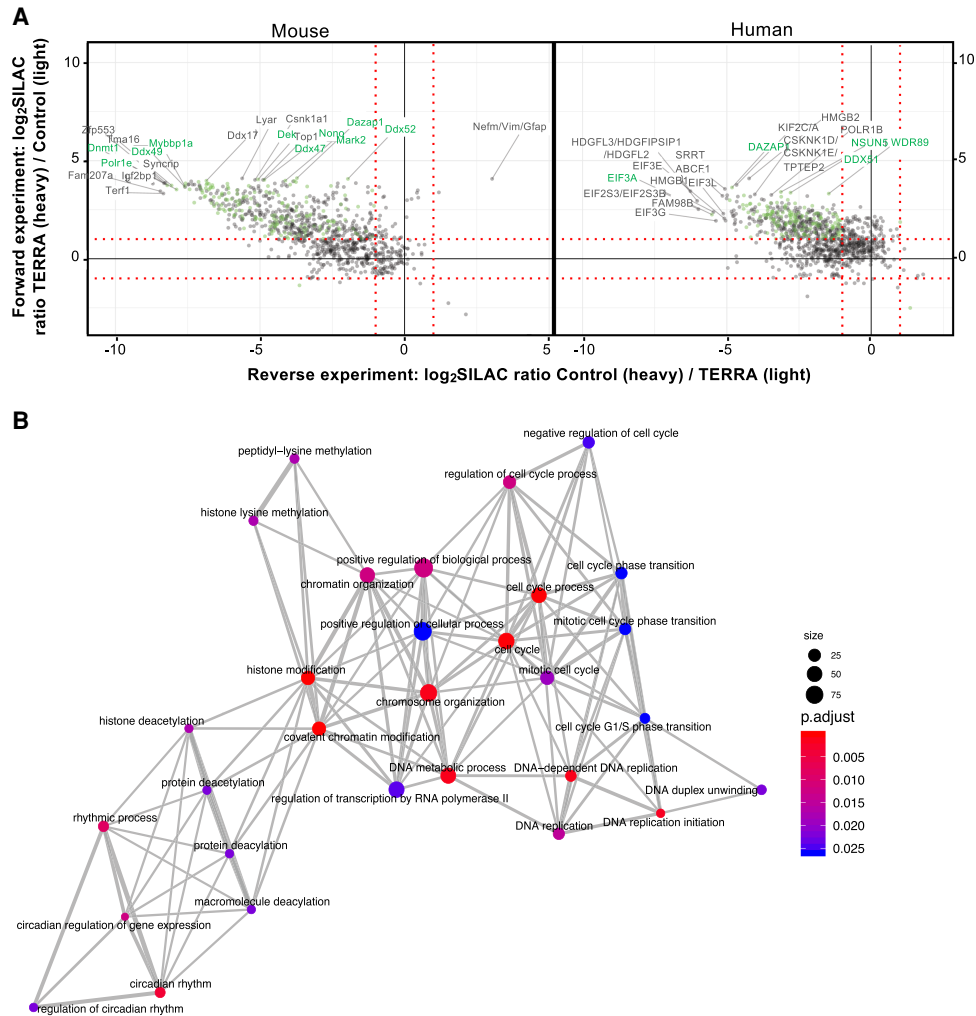


FIGURE 8. Comparison of human and mouse TERRA interactomes. (A) Two-dimensional interaction plot for the TERRA pull-down under 250 mM sodium chloride washing conditions both for murine and human (Scheibe et al. 2013) samples. Dotted red line indicates the enrichment threshold. The top TERRA enriched proteins are annotated, highlighting in green those that are shared between human and mouse data sets. (B) GO terms associated with the mouse 188 common interacting proteins were tested for overrepresentation (with Fisher's exact test) against GO terms associated to all detected proteins in the mouse data set. Nodes represent GO categories and edges show overlapping gene sets. Color scale shows the significance (BH adjusted *P*-values) of the overrepresentation, and node sizes are proportional to the number of genes per GO category.

to address the possible functions of Chr 2 TERRA in mouse. We also identified an additional intrachromosomal source of $(UUAGGG)_n$ -containing transcripts emanating from the *Tsix* locus—located within the X-inactivation center (Xic)—in mouse ES cells exclusively. Interactions between the Xic and the PAR loci were previously shown to be required for the initiation of X-chromosome inactivation in ES cells (Chu et al. 2017a). Interestingly, the reported CHIRT experiments suggested that the *Tsix* locus was bound not only by PAR-TERRA, but also by $(UUAGGG)_n$ -containing transcripts not related to PAR-TERRA (Chu et al. 2017a). We therefore anticipate that, in ES cells, the $(UUAGGG)_n$ -containing transcripts emanating from the *Tsix* locus may interact with PAR-TERRA to promote Xic:PAR pairing and X-chromosome inactivation.

The reason why mouse telomeres are mostly silent and Terf2 depletion is unable to de-repress telomere transcription is not clear. This may be related to the heterochromatic nature of mouse telomeres and subtelomeres that, unlike human telomeres, display high enrichment of H3K9me3 and H4K20me3 marks, together with abundant binding of Heterochromatin protein 1 (Blasco 2007; Rosenfeld et al. 2009). Alternatively, the lack of telomeric transcription may result from the absence, in the mouse genome, of dedicated subtelomeric promoters like in human cells. In line with this, production of the Telo 18q TERRA appears to result from the transcription of an upstream gene that initiates about 22 kb upstream of the first 18q telomeric repeats, and not from a subtelomeric promoter directly driving the transcription of telomeric repeats.

Despite their distinct genomic origins, increasing evidence suggests that TERRA molecules may play similar roles in human and mouse cells (Bettin et al. 2019). This sketches a fascinating picture where $(UUAGGG)_n$ -containing RNAs produced from different genomic loci may exert similar functions through their binding *in trans* to various genomic loci, including telomeres. Our comparative study of TERRA-interacting proteins, identified through SILAC-based pull-down screens in both species, reinforced the idea of common functions shared by human and mouse TERRA. Among the common interacting proteins, we found subunits of the PRC2 complex, Aurora kinases, proteins involved in DNA replication and repair, rRNA metabolism, pre-mRNA processing and Pol II-dependent transcription. Interestingly, interactions between Aurora kinase B (AURKB) and centromeric RNAs in mouse ES cells were previously reported to be involved in the recruitment of telomerase to chromosome ends in S phase (Mallm and Rippe 2015). AURKB was proposed to bind to centromeric RNAs before interacting with the telomerase complex to enhance its activity. It is therefore tempting to speculate that TERRA may similarly be implicated, through its interaction with AURKB, into the activation and/or recruitment of telomerase to telomeres, in line with previous observations in yeast (Cusanelli et al. 2013; Moravec et al. 2016). A more recent study reported on the localization of AurkB at mouse ESC telomeres where the kinase modulates Terf1 affinity for telomeres and participates in telomeric integrity (Chan et al. 2017). The interaction between AURKB and TERRA may participate in overall telomere protection through the modulation of shelterin affinity for telomeres. Common pulled-down proteins also included components of the paraspeckle subnuclear bodies and the Dazap1 RNA-binding protein. Interaction between TERRA and the human NONO paraspeckle component was recently confirmed and shown to suppress RNA:DNA hybrid-induced telomere instability; the same study confirmed the interaction between TERRA and the DAZAP1 RNA-binding protein (Petti et al. 2019). HNRNP proteins and the BLM helicase, previously identified as human TERRA interactors (Deng et al. 2009; Flynn et al. 2011; Petti et al. 2019), were also recovered in the screens and we validated the interaction between PAR-TERRA and mouse Blm by RIP. Interestingly, two classes of proteins appeared to be differentially recovered in the screens. The first one comprises centromeric proteins that were abundantly recovered in the mouse pull-down screen, suggesting a function for mouse TERRA in centromere assembly/stability. The second group includes proteins from the exosome complex (EXOSC) that were enriched in the human pull-down screen, but not in the mouse screen, suggesting possible distinct degradation machineries in human and mouse.

As we found a good overlap between our pull-down screen using mouse cell extracts and the *in vivo* iDRiP screen in ES cells (Chu et al. 2017b), in which PAR-

TERRA was likely the most abundant TERRA species authors were looking at, these results support the idea that, whether they originate from the transcription of telomeres or the PAR locus, these $UUAGGG$ -rich noncoding RNA molecules interact with the same protein sets and are likely to serve similar functions in the cell. Atrx, previously reported to interact with mouse TERRA (Chu et al. 2017b) was, however, not recovered in our pull-down screens using the $(UUAGGG)_8$ probe. This may be related to the observation that, although a 83-nt *in vitro* synthesized TERRA RNA probe was efficiently shifted by Atrx in electrophoretic mobility shift assays, a 30-nt TERRA was shifted to a much lesser degree (Chu et al. 2017b). This suggests that the oligonucleotide used in our studies might not be long enough to allow for efficient interaction with Atrx. Our biotinylated $(UUAGGG)_8$ probe similarly failed to pull-down other mouse iDRiP candidates, including Rtel1, Rpa1/2, Ctc1, Stn1, or Pml (Chu et al. 2017b), with important functions in telomere biology. This may be explained either by the size of the probe, as explained above, or by the need for a chromatin context that is not recapitulated in the *in vitro* pull-down assay. Alternatively, interactions may involve non- $(UUAGGG)_n$ sequences.

Although the good overlap with the *in vivo* iDRiP screen in mouse ES cells indicates that our pull-down screen efficiently recovered TERRA-interacting proteins, future experiments will be required to properly address TERRA functions in human and mouse cells using tools to specifically deplete TERRA or PAR-TERRA transcripts.

MATERIALS AND METHODS

Cell culture

The following mouse cell lines were used in this study: 3T3 spontaneously immortalized embryonic fibroblasts (ATCC), L929 spontaneously immortalized adult and adipose tissue-derived fibroblasts (ECACC), M1 myeloblast cell line (ECACC), RAW264.7 Abelson murine leukemia virus transformed macrophages (ATCC), Neuro-2a neuroblastoma cell line (ATCC) and J774A.1 reticulum cell sarcoma cell line (monocytes/macrophages) (ATCC) and were kindly provided by Thomas Michiels (de Duve Institute). Mouse embryonic fibroblasts (MEF) were prepared from CD1 mice using standard protocols and were kindly provided by Frédéric Lemaigre (de Duve Institute). ES cells were kindly provided by Olivier De Backer (Université de Namur). The human cells used in this study were previously described: HCA2 (Amout et al. 2010) and HFF2 (Mattiussi et al. 2012) human foreskin fibroblasts, HeLa cervix cancer cell line (ATCC), LB37 non-small cell lung cancer cell line (Tilman et al. 2009) and SW39 SV40T-immortalized fetal lung fibroblasts (Tilman et al. 2009, kindly provided by W. Wright, UT Southwestern Medical Center). Cells were cultured in EMEM (L929, Neuro-2a, SW39, HFF2, HCA2, LB37), DMEM (3T3, J774A.1, RAW264.7, MEF, HeLa), or RPMI (M1). All media were from Gibco and enriched with 1× Penicillin-Streptomycin (Gibco) and 10% FBS (Gibco). Cells were grown at

93. These are all strand-specific RNA-seq data sets with a high sequencing coverage (>140 million paired reads). Fastq files were all processed using the same pipeline. Read quality control was performed using FastQC v0.11.8 (Andrews 2010) and low-quality sequences were removed using Trimmomatic v0.38 (Bolger et al. 2014). Filtered reads were aligned on GRCm38 mouse genome using HISAT2 v2-2.1.0 (Kim et al. 2015). Gene expression levels were evaluated using featureCounts v2.0.0 (Liao et al. 2014) and Mus_musculus.GRCm38.94.gtf. Regions corresponding to Chr 2, PAR, Telo 18q, and Tsix PCR products were artificially introduced in the gtf file as fictive genes to allow their quantification. Strand specificity was taken into account and only unambiguously mapped reads were considered for quantification. Bam files were further converted to tdf files using igvtools v2.3.98 (Robinson et al. 2011) before viewing alignment data in the IGV browser.

SILAC labeling and nuclear extract preparation

Mouse R1/E embryonic stem cells (ATCC) were SILAC-labeled in DMEM (-Arg, -Lys) medium containing 42 mg/L $^{13}\text{C}_6^{15}\text{N}_4$ L-arginine (Euriso-Top) and 73 mg/L $^{13}\text{C}_6^{15}\text{N}_2$ L-lysine (Euriso-Top) or the corresponding concentration of unlabeled amino acids (Sigma-Aldrich). Medium was supplemented with 10% dialyzed fetal bovine serum (PAA), 1× nonessential amino acids (Gibco), 50 μM 2-mercaptoethanol (Gibco), 3 μM CT-99021 (Biomol), 1 μM PD-0325901 (Biomol), 1 mM sodium pyruvate, 100 U/mL LIF (Millipore), 100 U/mL penicillin and 100 $\mu\text{g}/\text{mL}$ streptomycin (Gibco). Nuclear extracts were prepared essentially as previously described (Scheibe et al. 2013) and were shock frozen in liquid nitrogen and stored at -80°C until use.

RNA pull-downs, mass spectrometry, and data analysis

RNA pull-downs with SILAC nuclear extracts were performed as previously described (Scheibe et al. 2013). Pull-downs were separated on a 4%–12% NOVEX gradient SDS gel (Life Technologies) for 50 min at 180 V in 1× MOPS buffer (Life Technologies) and processed into peptides as previously described (Scheibe et al. 2013). For mass spectrometry analyses, peptides were separated on a 20 cm self-packed column with 75- μm diameter filled with ReproSil-Pur 120 C₁₈-aq (Dr. Maisch GmbH) mounted to an EASY HPLC 1000 (Thermo Fisher) and sprayed online into a Q Exactive Plus mass spectrometer (Thermo Fisher). We used a 120-min gradient from 2% to 60% acetonitrile in formic acid at a flow of 225 nL/min. The mass spectrometer was operated with a top 10 MS/MS data-dependent acquisition scheme per MS full scan. Mass spectrometry raw data were searched using the Andromeda search engine (Cox et al. 2011) integrated into MaxQuant suite 1.5.2.8 (Cox and Mann 2008) using the ENSEMBL Mus_musculus.GRCm38 protein database (57,751 entries). The human data (Scheibe et al. 2013) was reanalyzed with MaxQuant suite 1.5.2.8 using the ENSEMBL Human GRCh38 protein database (102,915 entries). In both analyses, carbamidomethylation at cysteine was set as fixed modification while methionine oxidation and protein N-acetylation were considered as variable modifications. Match between run options was activated. Prior to bioinformatic analyses, reverse hits, proteins only identified by site, protein groups based on 1 unique peptide

and known contaminants were removed. Further filtering and graphical representation was done in an R framework incorporating the ggplot2 package (Wickham 2016; R Core Team 2017). To determine murine/human orthologs, we queried the Compara database provided within ENSEMBL (Zerbino et al. 2018) using the biomaRt software package (Durinck et al. 2005, 2009). Further, we queried the gene ontology (GO) database (Ashburner et al. 2000; The Gene Ontology Consortium 2017) for functional analysis. We tested our enriched proteins against all detected proteins using overrepresentation tests (FDR < 0.05) as implemented in the clusterProfiler package (Yu et al. 2012). Finally, functional associations among enriched proteins were highlighted using STRING (version 10.5) (Szklarczyk et al. 2017).

RNA immunoprecipitation (RIP)

Approximately 10^7 2575i MEF were cross-linked with 1% formaldehyde (Sigma-Aldrich) for 15 min at RT, before quenching with 125 mM glycine for 5 min at RT. After two washes with ice-cold PBS, cells were scraped from dishes and centrifuged at 1000g for 5 min at RT. Pellets were resuspended in 2 mL PBS, 2 mL Nuclear isolation buffer (40 mM Tris-HCl, pH 7.5, 20 mM MgCl₂, 4% Triton X-100, 1.28 M sucrose) and 6 mL water before incubation on ice for 20 min with frequent mixing. After centrifugation at 2500g for 15 min at 4°C, pellets were resuspended in 1 mL of RIP buffer (150 mM KCl, 25 mM Tris-HCl, pH 7.5, 5 mM EDTA, 0.5% Nonidet-P40, 1 mM DTT, 100 U/mL RNase Out [Thermo Fisher Scientific] and EDTA-free Protease Inhibitor Cocktail [Sigma-Aldrich]). Nuclear extracts were sonicated twice using a Bioruptor apparatus (Diagenode) before centrifugation at maximum speed for 10 min at 4°C. Sonicated extracts were precleared by incubation with Dynabeads Protein A (Thermo Fisher Scientific), preblocked with BSA and *E. coli* tRNA, on a rotating wheel for 1 h at 4°C. Precleared lysates were then diluted in RIP buffer to a final concentration of 1 mg/mL and 1 mL was used for each IP with 2 μg of the respective antibodies (rabbit anti-Blm [Bethyl Laboratories A300-110A], rabbit anti-Suz12 [Abcam ab12073] and rabbit anti-Terf2 [Novus Biologicals NB110-57130]). After 3 h incubation at 4°C on a rotating wheel, preblocked Dynabeads Protein A was added to the samples for an overnight incubation at 4°C. Beads were washed three times with RIP buffer and once with high salt RIP buffer (300 mM KCl) before elution with 25 mM Tris-HCl, pH 7.5, 5 mM EDTA, 0.5% SDS with 0.1 μg Proteinase K for 45 min at 45°C, and then overnight at 65°C to revert cross-link. Eluted RNA was purified using Nucleospin RNA kit (Macherey-Nagel) and cDNA was synthesized using Superscript IV (Thermo Fisher Scientific) and 10 μM Telo primers as described above.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

We are grateful to Titia de Lange (The Rockefeller University, New York, USA) for Terf1 antibody, to Thomas Michiels (de Duve Institute, UCLouvain, Belgium) for mouse cell lines, to Frédéric Lemaigre (de Duve Institute, Brussels, Belgium) for MEF, to Eros

Lazzerini Denchi (The Scripps Research Institute, La Jolla, California, USA) for CreERT2 *Terf2* F/+ and F/F MEF, and to Olivier De Backer (Université de Namur, Namur, Belgium) for the kind gift of mouse ES cells. We are most grateful to Amandine Van Beneden for help with RNA slot-blot experiments. N.V. and A.L. were supported by a grant from the D.G. Higher Education and Scientific Research of the French Community of Belgium (Actions de Recherche Concertées). A.D. is a recipient from the Fonds National de la Recherche Scientifique (FNRS). Work in the Azzalin laboratory was supported by the European Molecular Biology Organization (IG3576) and the Fundação para a Ciência e a Tecnologia (IF/01269/2015; PDTC/MED-ONC/28282/2017; PDTC/BIA-MOL/29352/2017). P.L.A. is the recipient of an FCT PhD fellowship (PD/BD/128284/2017). We are most grateful to the de Duve Institute for constant support.

Received May 6, 2020; accepted October 28, 2020.

REFERENCES

- Andrews S. 2010. *FastQC: a quality control tool for high throughput sequence data*. <http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>
- Arnoult N, Schluth-Bolard C, Letessier A, Drascovic I, Bouarich-Bourimi R, Campisi J, Kim S-H, Boussouar A, Ottaviani A, Magdinier F, et al. 2010. Replication timing of human telomeres is chromosome arm-specific, influenced by subtelomeric structures and connected to nuclear localization. *PLoS Genet* **6**: e1000920. doi:10.1371/journal.pgen.1000920
- Arnoult N, Van Beneden A, Decottignies A. 2012. Telomere length regulates TERRA levels through increased trimethylation of telomeric H3K9 and HP1 α . *Nat Struct Mol Biol* **19**: 948–956. doi:10.1038/nsmb.2364
- Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**: 25–29. doi:10.1038/75556
- Azzalin CM, Reichenbach P, Khoraiuli L, Giulotto E, Lingner E. 2007. Telomeric repeat-containing RNA and RNA surveillance factors at mammalian chromosome ends. *Science* **318**: 798–801. doi:10.1126/science.1147182
- Bettin N, Oss Pegorar C, Cusanelli E. 2019. The emerging roles of TERRA in telomere maintenance and genome stability. *Cells* **8**: 246. doi:10.3390/cells8030246
- Biffi G, Tannahill D, Balasubramanian S. 2012. An intramolecular G-quadruplex structure is required for binding of telomeric repeat-containing RNA to the telomeric protein TRF2. *J Am Chem Soc* **134**: 11974–11976. doi:10.1021/ja305734x
- Blasco MA. 2007. The epigenetic regulation of mammalian telomeres. *Nat Rev Genet* **8**: 299–309. doi:10.1038/nrg2047
- Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120. doi:10.1093/bioinformatics/btu170
- Caslini C, Connelly JA, Sema A, Broccoli D, Hess JL. 2009. MLL associates with telomeres and regulates telomeric repeat-containing RNA transcription. *Mol Cell Biol* **29**: 4519–4526. doi:10.1128/MCB.00195-09
- Chan FL, Vinod B, Novy K, Schittenhelm RB, Huang C, Udugama M, Nunez-Iglesias J, Lin JI, Hii L, Chan J, et al. 2017. Aurora kinase B, a novel regulator of TERF1 binding and telomeric integrity. *Nucleic Acids Res* **45**: 21. doi:10.1093/nar/gkx904
- Chu HP, Froberg JE, Kesner B, Oh HJ, Ji F, Sadreyev R, Pinter SF, Lee JT. 2017a. PAR-TERRA directs homologous sex chromosome pairing. *Nat Struct Mol Biol* **24**: 620–631. doi:10.1038/nsmb.3432
- Chu HP, Cifuentes-Rojas C, Kesner B, Aeby E, Lee HG, Wei C, Oh HJ, Boukhali M, Haas W, Lee JT. 2017b. TERRA RNA antagonizes ATRX and protects telomeres. *Cell* **170**: 86–101. doi:10.1016/j.cell.2017.06.017
- Cox J, Mann M. 2008. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**: 1367–1372. doi:10.1038/nbt.1511
- Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. 2011. Andromeda: a peptide search engine integrated into the MaxQuant environment. *J Proteome Res* **10**: 1794–1805. doi:10.1021/pr101065j
- Cusanelli E, Romero CA, Chartrand P. 2013. Telomeric noncoding RNA TERRA is induced by telomere shortening to nucleate telomerase molecules at short telomeres. *Mol Cell* **51**: 780–791. doi:10.1016/j.molcel.2013.08.029
- Deng Z, Norseen J, Wiedmer A, Riethman H, Lieberman PM. 2009. TERRA RNA binding to TRF2 facilitates heterochromatin formation and ORC recruitment at telomeres. *Mol Cell* **35**: 403–413. doi:10.1016/j.molcel.2009.06.025
- Deng Z, Wang Z, Xiang C, Molczan A, Baubet V, Conejo-Garcia J, Xu X, Lieberman PM, Dahmane N. 2012. Formation of telomeric repeat-containing RNA (TERRA) foci in highly proliferating mouse cerebellar neuronal progenitors and medulloblastoma. *J Cell Sci* **125**: 4383–4394. doi:10.1242/jcs.108118
- Diman A, Decottignies A. 2018. Genomic origin and nuclear localization of TERRA telomeric repeat-containing RNA: from darkness to dawn. *FEBS J* **285**: 1389–1398. doi:10.1111/febs.14363
- Diman A, Boros J, Poulain F, Rodriguez J, Pumelle M, Episkopou H, Bertrand L, Francaux M, Deldicque L, Decottignies A. 2016. Nuclear respiratory factor 1 and endurance exercise promote human telomere transcription. *Sci Adv* **2**: e1600031. doi:10.1126/sciadv.1600031
- Durinck S, Moreau Y, Kasprzyk A, David S, De Moor B, Brazma A, Huber W. 2005. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **21**: 3439–3440. doi:10.1093/bioinformatics/bti525
- Durinck S, Spellman PT, Birney E, Huber W. 2009. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc* **4**: 698–705. doi:10.1038/nprot.2009.97
- Feretziaki M, Lingner J. 2017. A practical qPCR approach to detect TERRA, the elusive telomeric repeat-containing RNA. *Methods* **114**: 39–45. doi:10.1016/j.ymeth.2016.08.004
- Feretziaki M, Renck Nunes P, Lingner J. 2019. Expression and differential regulation of human TERRA at several chromosome ends. *RNA* **25**: 1470–1480. doi:10.1261/rna.072322.119
- Flynn RL, Centore RC, O’Sullivan RJ, Rai R, Tse A, Songyang Z, Chang S, Karlseder J, Zou L. 2011. TERRA and hnRNPA1 orchestrate an RPA-to-POT1 switch on telomeric single-stranded DNA. *Nature* **471**: 532–536. doi:10.1038/nature09772
- The Gene Ontology Consortium. 2017. Expansion of the gene ontology knowledgebase and resources. *Nucleic Acids Res* **45**: D331–D338. doi:10.1093/nar/gkw1108
- Göhring J, Fulcher N, Jacak J, Riha K. 2014. TeloTool: a new tool for telomere length measurement from terminal restriction fragment analysis with improved probe sensitivity correction. *Nucleic Acids Res* **42**: e21. doi:10.1093/nar/gkt1315
- Hesselberth JR. 2013. Lives that introns lead after splicing. *WIREs RNA* **4**: 677–691. doi:10.1002/wrna.1187
- Hirashima K, Seimiya H. 2015. Telomeric repeat-containing RNA/G-quadruplex-forming sequences cause genome-wide alteration of

- gene expression in human cancer cells *in vivo*. *Nucleic Acids Res* **43**: 2022–2032. doi:10.1093/nar/gkv063
- Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**: 357–360. doi:10.1038/nmeth.3317
- Koskas S, Decottignies A, Dufour S, Pezet M, Verdel A, Vourc'h C, Faure V. 2017. Heat shock factor 1 promotes TERRA transcription and telomere protection upon heat stress. *Nucleic Acids Res* **45**: 6321–6333. doi:10.1093/nar/gkx208
- Le Berre G, Hossard V, Riou JF, Guieysse-Peugeot AL. 2019. Repression of TERRA expression by subtelomeric DNA methylation is dependent on NRF1 binding. *Int J Mol Sci* **20**: 2791. doi:10.3390/ijms20112791
- Lee YW, Arora R, Wischniewski H, Azzalin CM. 2018. TRF1 participates in chromosome end protection by averting TRF2-dependent telomeric R loops. *Nat Struct Mol Biol* **25**: 147–153. doi:10.1038/s41594-017-0021-5
- Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**: 923–930. doi:10.1093/bioinformatics/btt656
- López de Silanes I, Stagno d'Alcontres M, Blasco MA. 2010. TERRA transcripts are bound by a complex array of RNA-binding proteins. *Nat Commun* **1**: 33. doi:10.1038/ncomms1032
- López de Silanes I, Graña-Castro O, Luigia De Bonis M, Dominguez O, Pisano DG, Blasco MA. 2014. Identification of TERRA locus unveils a telomere protection role through association to nearly all chromosomes. *Nat Commun* **5**: 4723. doi:10.1038/ncomms5723
- Mallm JP, Rippe K. 2015. Aurora kinase B regulates telomerase activity via a centromeric RNA in stem cells. *Cell Rep* **11**: 1667–1678. doi:10.1016/j.cellrep.2015.05.015
- Marión RM, Montero JJ, de Silanes I L, Graña-Castro O, Martínez P, Schoeftner S, Palacios-Fábrega JA, Blasco MA. 2019. TERRA regulate the transcriptional landscape of pluripotent cells through TRF1-dependent recruitment of PRC2. *Elife* **8**: e44656. doi:10.7554/eLife.44656
- Mattiussi M, Tilman G, Lenglez S, Decottignies A. 2012. Human telomerase represses ROS-dependent cellular responses to Tumor Necrosis Factor- α without affecting NF- κ B activation. *Cell Signal* **24**: 708–717. doi:10.1016/j.cellsig.2011.11.004
- Mayr C. 2017. Regulation by 3'-untranslated regions. *Ann Rev Genet* **51**: 171–194. doi:10.1146/annurev-genet-120116-024704
- Moravec M, Wischniewski H, Bah A, Hu Y, Liu N, Lafranchi L, King MC, Azzalin CM. 2016. TERRA promotes telomerase-mediated telomere elongation in *Schizosaccharomyces pombe*. *EMBO Rep* **17**: 999–1012. doi:10.15252/embr.201541708
- Nergadze SG, Famung BO, Wischniewski H, Khoraiuli L, Vitelli V, Chawla R, Giulotto E, Azzalin CM. 2009. CpG-island promoters drive transcription of human telomeres. *RNA* **15**: 2186–2194. doi:10.1261/ma.1748309
- Okamoto K, Bartocci C, Ouzounov I, Diedrich JK, Yates JR, Lazzerini Denchi E. 2013. A two-step mechanism for TRF2-mediated chromosome-end protection. *Nature* **494**: 502–505. doi:10.1038/nature11873
- Petti E, Buemi V, Zappone A, Schillaci O, Veneziano Broccia P, Dinami R, Matteoni S, Benetti R, Schoeftner S. 2019. SFPQ and NONO suppress RNA:DNA-hybrid-related telomere instability. *Nat Commun* **10**: 1001. doi:10.1038/s41467-019-08863-1
- Porro A, Feuerhahn S, Delafontaine J, Riethman H, Rougemont J, Lingner J. 2014a. Functional characterization of the TERRA transcriptome at damaged telomeres. *Nat Commun* **5**: 5379. doi:10.1038/ncomms6379
- Porro A, Feuerhahn S, Lingner J. 2014b. TERRA-reinforced association of LSD1 with MRE11 promotes processing of uncapped telomeres. *Cell Rep* **6**: 765–776. doi:10.1016/j.cellrep.2014.01.022
- R Core Team. 2017. *R: a language and environment for statistical computing*. <https://www.R-project.org/>
- Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nat Biotechnol* **29**: 24–26. doi:10.1038/nbt.1754
- Rosenfeld JA, Wang Z, Schones DE, Zhao K, DeSalle R, Zhang MQ. 2009. Determination of enriched histone modifications in non-genic portions of the human genome. *BMC Genomics* **10**: 143. doi:10.1186/1471-2164-10-143
- Rudenko G, Van der Ploeg LH. 1989. Transcription of telomere repeats in protozoa. *EMBO J* **8**: 2633–2638.
- Sagie S, Toubiana S, Hartono SR, Katzir H, Tzur-Gilat A, Havazelet S, Francastel C, Velasco G, Chédin F, Selig S. 2017. Telomeres in ICF syndrome cells are vulnerable to DNA damage due to elevated DNA:RNA hybrids. *Nat Commun* **8**: 14015. doi:10.1038/ncomms14015
- Scheibe M, Arnoult N, Kappei D, Buchholz F, Decottignies A, Butter F, Mann M. 2013. Quantitative interaction screen of telomeric repeat-containing RNA reveals novel TERRA regulators. *Genome Res* **23**: 2149–2157. doi:10.1101/gr.151878.112
- Schoeftner S, Blasco MA. 2008. Developmentally regulated transcription of mammalian telomeres by DNA-dependent RNA polymerase II. *Nat Cell Biol* **10**: 228–236. doi:10.1038/ncb1685
- Schoeftner S, Blanco R, Lopez de Silanes I, Muñoz P, Gómez-López G, Flores JM, Blasco MA. 2009. Telomere shortening relaxes X chromosome inactivation and forces global transcriptome alterations. *Proc Natl Acad Sci* **106**: 19393–19398. doi:10.1073/pnas.0909265106
- Solovei I, Gaginskaya ER, Macgregor HC. 1994. The arrangement and transcription of telomere DNA sequences at the ends of lampbrush chromosomes of birds. *Chromosome Res* **2**: 460–470. doi:10.1007/bf01552869
- Szklarczyk D, Morris JH, Cook H, Kuhn M, Wyder S, Simonovic M, Santos A, Doncheva NT, Roth A, Bork P, et al. 2017. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res* **45**: D362–D368. doi:10.1093/nar/gkw937
- Tilman G, Liorot A, Van Beneden A, Arnoult N, Londoño-Vallejo JA, De Smet C, Decottignies A. 2009. Subtelomeric DNA hypomethylation is not required for telomeric sister chromatid exchanges in ALT cells. *Oncogene* **28**: 1682–1693. doi:10.1038/onc.2009.23
- Van Beneden A, Arnoult N, Decottignies A. 2013. Telomeric RNA expression: length matters. *Front Oncol* **3**: 178. doi:10.3389/fonc.2013.00178
- Viceconte N, Dheur MS, Majerova E, Pierreux CE, Baurain JF, van Baren N, Decottignies A. 2017. Highly aggressive metastatic melanoma cells unable to maintain telomere length. *Cell Rep* **19**: 2529–2543. doi:10.1016/j.celrep.2017.05.046
- Wickham H. 2016. *ggplot2: elegant graphics for data analysis*. Springer-Verlag, New York.
- Yehezkel S, Segev Y, Viegas-Péquignot E, Skorecki K, Selig S. 2008. Hypomethylation of subtelomeric regions in ICF syndrome is associated with abnormally short telomeres and enhanced transcription from telomeric regions. *Hum Mol Genet* **17**: 2776–2789. doi:10.1093/hmg/ddn177
- Yu G, Wang LG, Han Y, He QY. 2012. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**: 284–287. doi:10.1089/omi.2011.0118
- Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, Billis K, Cummins C, Gall A, Girón CG, et al. 2018. Ensembl 2018. *Nucleic Acids Res* **46**: D754–D761. doi:10.1093/nar/gkx1098
- Zhang L-F, Ogawa Y, Ahn JY, Namekawa SH, Silva SS, Lee JT. 2009. Telomeric RNAs mark sex chromosomes in stem cells. *Genetics* **182**: 685–692. doi:10.1534/genetics.109.103093



RNA

A PUBLICATION OF THE RNA SOCIETY

PAR-TERRA is the main contributor to telomeric repeat-containing RNA transcripts in normal and cancer mouse cells

Nikenza Viceconte, Axelle Lorient, Patrícia Lona Abreu, et al.

RNA 2021 27: 106-121 originally published online October 30, 2020
Access the most recent version at doi:[10.1261/ma.076281.120](https://doi.org/10.1261/ma.076281.120)

Supplemental Material

<http://rnajournal.cshlp.org/content/suppl/2020/10/30/rna.076281.120.DC1>

References

This article cites 61 articles, 10 of which can be accessed free at:
<http://rnajournal.cshlp.org/content/27/1/106.full.html#ref-list-1>

Creative Commons License

This article is distributed exclusively by the RNA Society for the first 12 months after the full-issue publication date (see <http://rnajournal.cshlp.org/site/misc/terms.xhtml>). After 12 months, it is available under a Creative Commons License (Attribution-NonCommercial 4.0 International), as described at <http://creativecommons.org/licenses/by-nc/4.0/>.

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

Use CRISPRmod for targeted modulation of endogenous gene expression to validate siRNA data

horizon
a PerkinElmer company

To subscribe to *RNA* go to:
<http://rnajournal.cshlp.org/subscriptions>

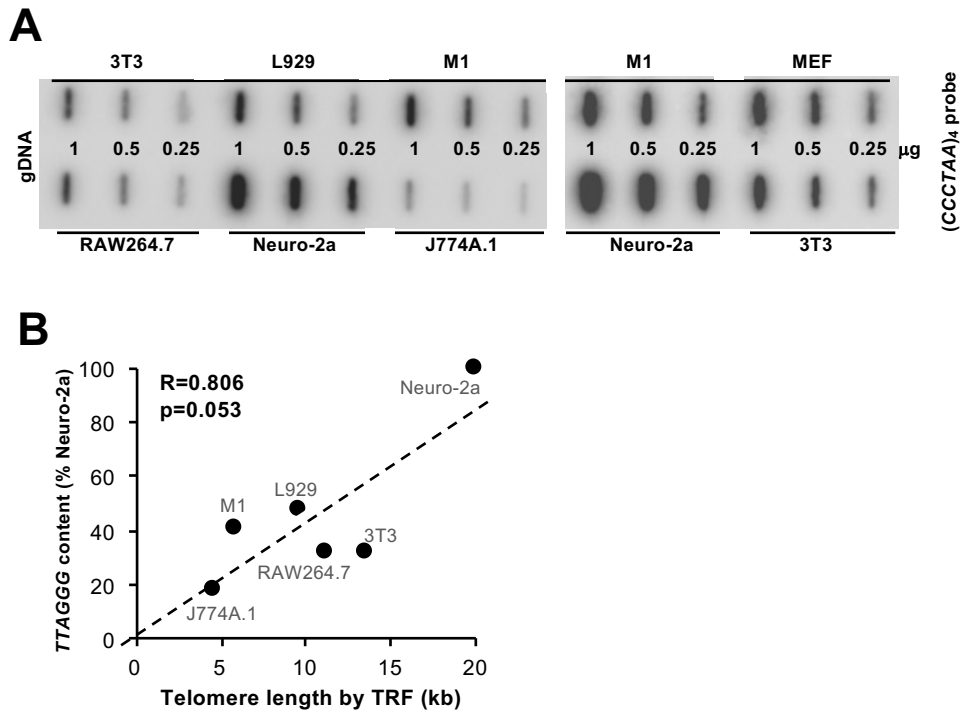


Figure S1. A. Evaluation of total *TTAGGG* content in mouse cell lines using slot-blot. Genomic DNA (gDNA) was isolated from the indicated cell lines and 1, 0.5 or 0.25 μg were blotted and hybridized with a radioactive $(\text{CCCTAA})_4$ probe. **B.** Total *TTAGGG* content (normalized to Neuro-2a cells) was plotted against telomere length evaluated by TRF (see Figure 2A).

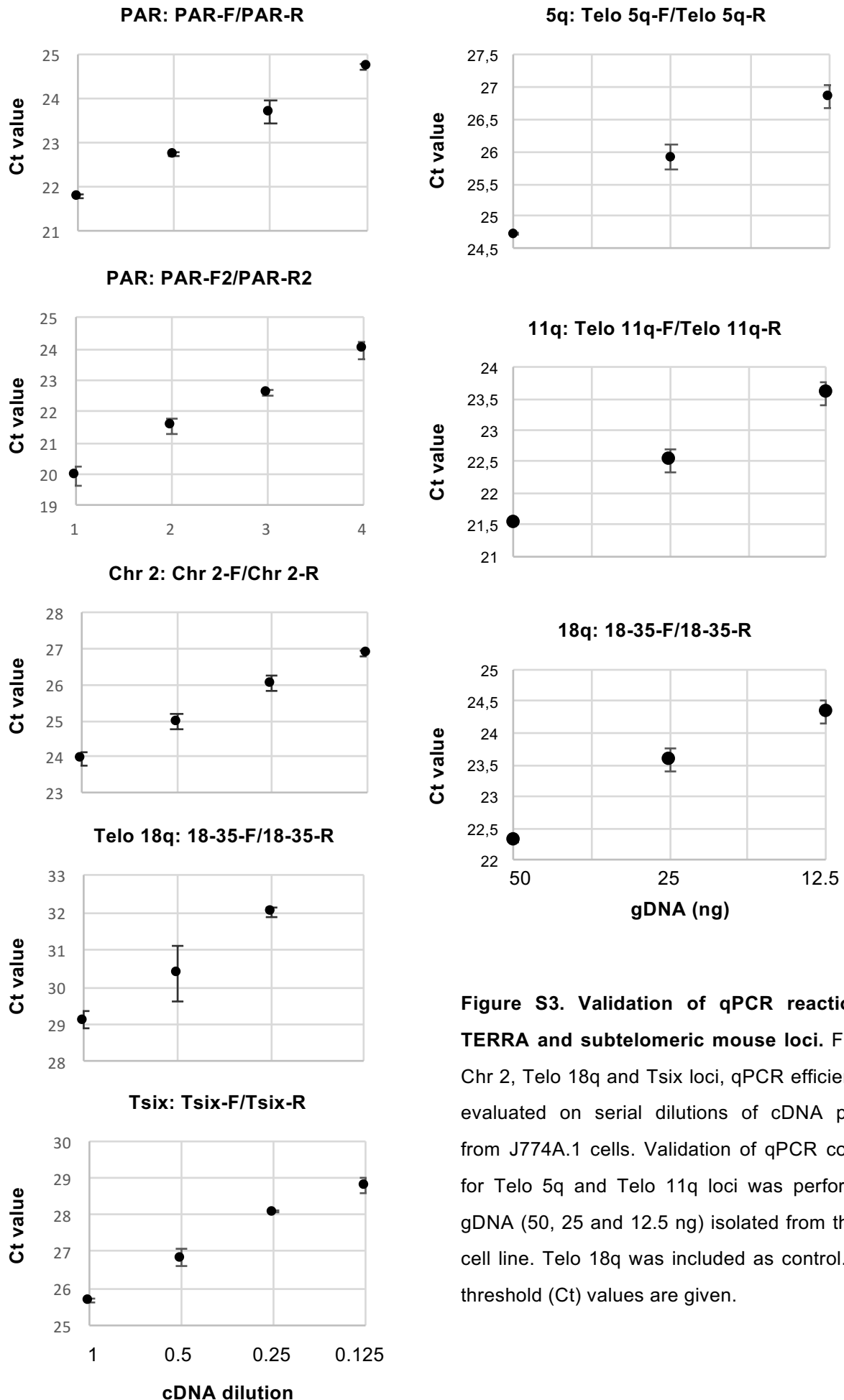


Figure S3. Validation of qPCR reactions for TERRA and subtelomeric mouse loci. For PAR, Chr 2, Telo 18q and Tsix loci, qPCR efficiency was evaluated on serial dilutions of cDNA prepared from J774A.1 cells. Validation of qPCR conditions for Telo 5q and Telo 11q loci was performed on gDNA (50, 25 and 12.5 ng) isolated from the same cell line. Telo 18q was included as control. Critical threshold (Ct) values are given.

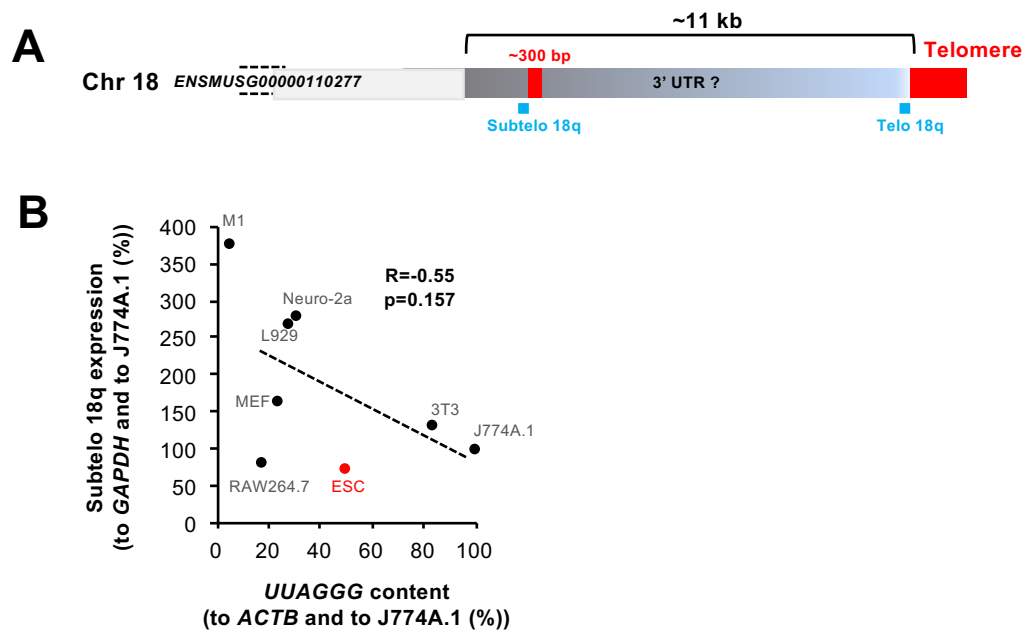


Figure S4. Internal telomeric repeats at 18q subtelomere do not contribute to the bulk of *UUAGGG* repeats in mouse cells. A. Overview of 18q chromosome end. Red boxes indicate *TTAGGG*-rich loci. Position of PCR products is indicated by light blue boxes. **B.** Expression of 18q subtelomere was evaluated by qRT-PCR, normalized to *GAPDH* and to *J774A.1* cells and plotted against the relative *UUAGGG* content of cells lines, measured by slot-blot and normalized to *ACTB* and to *J774A.1* cells.

A

Aatf	Cenpv	Ddx47	Fam207a	Ilf2	Lyar	Nono	Poldip3	Ran Rasl2-9	Rps14	Srfbp1	U2af1 U2af14	Zfp646
Aen	Chd1	Ddx49	Fam32a	Ilf3	Mark2	Nop2	Polr1d	Rbbp6	Rps25	Srrm1	U2af2	Zfp689
Alkbh5	Chd3	Ddx5	Fam60a	Imp3	Matr3 Gm28285	Npm1	Polr1e	Rbm10	Rps5	Srsf3 Gm12355	Ubtf	Zfy1 Zfp711 Zfx Zfy2
Alyref	Chd4	Ddx51	Fbxl6	Imp4	Mbd4	Npm3	Polr3a Polr1a	Rbm12b2	Rps7 Gm9493	Ssbp1	Uhrf1 Uhrf2	Zfyve26 Urb2
Atad5	Chd6	Ddx55	Frg1	Incenp	Mcm5	Nr0b1	Pop7	Rbm14 Gm21992	Rrp12	Ssrp1	Usp36	Zik1
Atp5a1	Chd8	Ddx56	Fus	Jarid2	Mcrs1	Nsd1	Ppib	Rbm15	Rrp1b	Sub1	Usp39	Znhit3
Atp5c1	Chd9	Dek	Fxr1	Kat8	Men1	Nsrp1	Ppih Gm7879	Rbm15b	Rrp8	Supt16	Usp7	Zscan26
Aurka	Clk3	Dhx33	Gdf3	Kdm2b 2a	Mki67	Nsun2	Ppil4	Rbm26	Rrp9	Suz12	Utf1	
Aurkb	Cmas	Dhx38	Glyr1	Kif18a	Morf4I2	Nsun5	Ppp1cc	Rbm27	Rsbn1	Syncrip	Utp18	
Aurkc	Coil	Dhx8	Gm17067	Kif23	Mphosph10	Numa1	Pqbp1	Rbm3	Rtcb	Taf1	Utp3	
Baz1a	Csnk1a1	Dimt1	Gm340	Klf2	Mrpl4	Nusap1	Prdm15	Rbm39	Rtraf	Taf10	Wdr18	
Baz1b	Csnk1d	Dmap1	Gm38394 Zbed6	Klf4	Msh6	Nvl	Prpf19	Rbmx2	Sart1	Taf15	Wdr46	
Baz2a	Csnk1e Csnk1d	Dnaja1	Gni3l	Knop1	Msl3	Nxf1	Prpf31	Rest	Scaf11	Taf4	Wdr89	
Bcas2	Cwc22 Gm13691 Gm13697 Gm13695 Gm13694 Gm13696 Gm13693 Gm13698	Dnd1	Gpatch4	Kpna2 Gm10184	Mtf2	Nxt1	Prpf38a	Rfc1	Sdad1	Taf5	Yeats4	
Blm	Cwc25	Dnmt1	Gtf3c1	Kri1	Mybbp1a	Orc1	Prpf4	Rfc3	Senp2 Gm5415	Taf7	Zcchc9	
Bud13	Cwf19I2	Drosha	Gtf3c4	Krr1	Nat10	Orc2	Prpf40a	Rfc4	Sfpq Gigyf2	Tardbp	Zfc3h1	
Caap1	Dazap1	Dynl1 Dylnl2 BC048507	Hnrnpab	Las1l	Ncl	Papd4	Prpf4b	Rnf2	Sgo2a Sgo2b	Terf1	Zfp1	
Cbx5	Ddx1	Eed	Hnrnpd	Lbr	Nefm Vim Gfap	Parp1	Prrc2c	Rp9	Skp1a	Tex10	Zfp260	
Cdc5l	Ddx17	Ehmt1	Hnrnpf	Leng8	Nip7	Patz1	Psip1	Rpl22	Slc25a5 Slc25a31	Tma16	Zfp281	
Cdk1	Ddx18	Elavl1	Hnrnph1	Lig3	Nipbl	Pcm1	Pspc1	Rpl22l1	Smarca4	Tmpo	Zfp292	
Cdk12 Cdk4 Cdk18 Cdk14 Cdk16 Cdk6 Cdk17 Cdk15 Cdk3-ps	Ddx28	Eloc	Hnrnph2	Llgl2	Nipsnap1	Pes1	Psrc1	Rpl23	Smarca5	Top1	Zfp37	
Cenpo	Ddx31	Emg1	Hnrnpl	Lsm4	Nob1	Pick1 Map4k4	Ptbp1	Rpl27 Rpl27-ps3	Snip1	Top2a	Zfp383	
Genpp	Ddx3x	Esco2	Hnrnpr	Luc7l	Nodal	Pip4k2c	Ptbp3	Rpl30	Son	Tpx2	Zfp462	
Genpq	Ddx4 Ddx3y	Esrrb Esrra Esrrg	Igf2bp1	Luc7l2	Nol10	Plrg1	Ptk6 Cdk11b	Rpl9 Rpl9-ps6 Rpl9-ps1	Spout1	Trip12	Zfp512	
Cenpu	Ddx46	Ewsr1	Ik	Luc7l3	Nom1	Pno1	Racgap1	Rps10	Srbd1	Tufm	Zfp553	

B

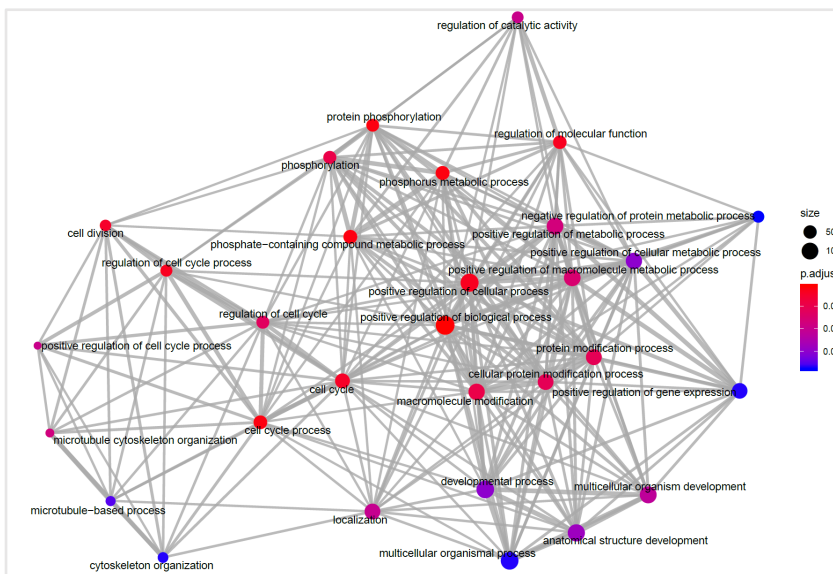


Figure S5. TERRA-enriched proteins in the mouse dataset. A. List of the 307 candidate proteins showing at least 4-fold enrichment. **B.** GO terms associated with the 307 proteins were tested for over-representation (with Fisher's exact test) against GO terms associated with all detected proteins (972). In order to highlight functional associations, the results are depicted as an enrichment map network where nodes represent GO categories and edges show overlapping gene sets. Color scale shows the significance (BH adjusted p-values) of the over-representation while node sizes are correlated with the number of genes found per GO category.

A

AC091959.3	CHD8	DDX56	HDAC2	MBD4	NSUN5	PPIH	RFC4	SERBP1	TJP2
AHCTF1	CHD9	DEK	HNRNPD	MCM2	NUMA1	PPP1CC	RFC5	SFPQ	TMPO
ALYREF	CMAS	DXH16	HNRNPH1	MCM4	NUP160	PRPF19	RIF1	SMARCA4	TOP2A
ATP5F1A	COIL	DMAP1	HNRNPL	MCM5	NXF1	PRPF4	RIOX2	SMARCC1	TOP2B
ATP5F1C	CSNK1D	DNAJA2	HSPA9	MCM6	NXT1	PRPF6	RNF2	SMARCD1	TPTEP2-CSNK1E
AURKA	CSNK1E	DNMT1	KDM2B	MCM7	ORC1	PSIP1	RPL23	SMU1	TRIP12
AURKB	DAZAP1	DSP	KHDRBS1	MEN1	ORC4	PSPC1	RPL9	SNRNP40	TUFM
AURKC	DDX1	DYNLL1	KHDRBS2	MKI67	ORC5	PTBP1	RPS14	SNRPA1	UHRF1
BAZ1A	DDX17	EED	KHDRBS3	MRPS30	PAK1IP1	PTBP3	RPS7	SNRPB2	URB1
BAZ2A	DDX24	EHMT1	KHSRP	MSH2	PATZ1	RACGAP1	RRP12	SNW1	URB2
BCAS2	DDX3X	EIF3A	KIF23	MTA1	PBRM1	RANBP2	RRP1B	SPTBN1	USP39
BUB3	DDX3Y	EMG1	KMT2A	MTA2	PCID2	RBBP5	RRP8	SRFBP1	USP7
BUD31	DDX4	EWSR1	KRI1	MYBBP1A	PES1	RBBP7	RRP9	SSRP1	UTP18
CD2BP2	DDX47	FXR1	LBR	NAT10	PLRG1	RBM10	RSBN1L	SUB1	WDR5
CDC40	DDX49	GATAD2A	LENG8	NOC4L	PNO1	RBM27	RTCB	SUPT16H	WDR82
CDC5L	DDX5	GIGYF2	LIG3	NOLC1	POLDIP3	RCC2	RTRAF	SUZ12	WDR89
CHD3	DDX51	GTF3C1	LRWD1	NONO	POLR1A	RFC1	RUVBL1	TECR	ZFC3H1
CHD4	DDX52	GTF3C4	MARK2	NOP14	POLR1E	RFC2	RUVBL2	THOC2	
CHD6	DDX55	HDAC1	MATR3	NSRP1	PPIB	RFC3	SART1	TJP1	

B

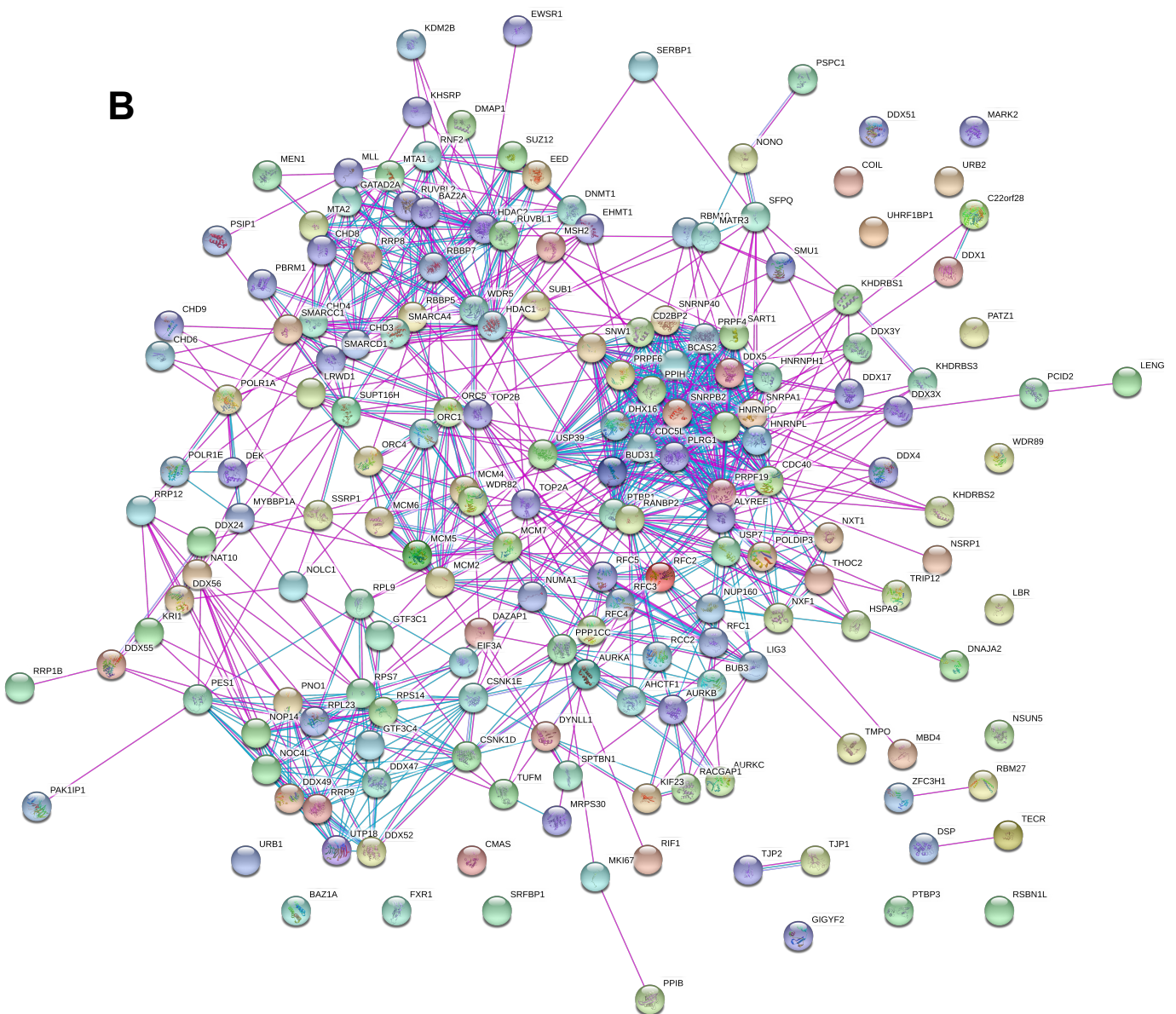


Figure S6. Common TERRA interacting proteins in human and mouse datasets. A. List of the 188 common candidate proteins in the SILAC pull-down screens (at least 2-fold enrichment). **B.** STRING analysis of the common interacting proteins.

3.2 RNA-dependent interactome allows network-based assignment of RNA-binding protein function

3.2.1 Summary

This project focused on function assignment of RBPs in *S. cerevisiae*. An AP-MS screen was designed to identify and quantify the interaction partners of 40 selected RBPs. These selections were based on their involvement in RNA pathways, in a way that several stages of an mRNA's life cycle would be covered. Hence, we selected RBPs involved in mRNA pre-processing, such as splicing, export from the nucleus, transport, localization and degradation. The selected RBPs were immunoprecipitated, and their PPI and RDI partners were quantified. Then, we proceeded with functional data analysis.

Protein domain enrichment revealed an overrepresentation of canonical RNA binding domains (RBDs) among the RDI group of several RBPs. Similarly, an overrepresentation of RNA binding-related GO molecular function terms was found within the RDI group for multiple RBPs. On the other hand, canonical RBDs and RNA binding-related terms were less predominant among the PPI group. Finally, a KEGG functional analysis was performed. It revealed an overrepresentation of RNA pathways in both the RDI and PPI groups, which, in most cases, overlapped with the RBP selection criteria. Additionally, we also identified involvement of such interactors in metabolic and synthesis pathways. Further functional implications for a subset of non-essential RBPs, extracted from the 40 immunoprecipitated RBPs, were explored with a knockout (KO) proteome screen.

The collected data were combined with manually curated *S. cerevisiae* protein complex annotations to create function-based networks. Sub-networks for splicing, export, ribosome, synthesis, metabolism, and degradation functionalities were created. Network-based assignment of RBP function was suggested based on the observed concurrent RBP binding patterns. RNA splicing was used as an exemplary process and thus further developed, revealing novel RBP associations with spliceosome protein complexes.

Overall, we provided an extensive RBP interactome network, systematically identifying RDIs and PPIs. Concurrent binding patterns within the network allowed us to suggest functional associations for the selected RBPs and their interaction partners. The resource is collected within the RBP interactome network explorer (RINE), which is available through an online interactive platform at <https://www.butterlab.org/RINE>.

3.2.2 Zusammenfassung

Dieses Projekt untersuchte die Funktionszuordnung von RBPs in *S. cerevisiae*. Ein AP-MS Ansatz wurde durchgeführt, um Interaktionspartner von 40 ausgewählten RBPs zu identifizieren und zu quantifizieren. Diese wurden aufgrund ihrer Beteiligung an RNA Prozessierungsschritten ausgewählt, um verschiedene Stadien des mRNA-Lebenszyklus abzudecken. Somit wurden RBPs ausgewählt, die an der mRNA-Vorprozessierung beteiligt sind, einschließlich dem Spleißen, Export aus dem Zellkern, Transport, Lokalisierung und Abbau. Die ausgewählten RBPs wurden immunpräzipitiert um ihre PPI- und RDI-Interaktionspartner zu quantifizieren. Anschließend führten wir eine funktionale Datenanalyse durch.

Die Anreicherung von Proteindomänen zeigte eine Überrepräsentation kanonischer RNA-Bindungsdomänen (RBDs) innerhalb der RDI-Gruppe mehrerer RBPs. Ebenso wurde eine Überrepräsentation von GO-Begriffen zu molekularen Funktionen im Zusammenhang mit RNA-Bindungen in der RDI-Gruppe für mehrere RBPs festgestellt. Demgegenüber waren kanonische RBDs und mit RNA-Bindung zusammenhängende Begriffe in der RBPs PPI-Gruppe weniger vorherrschend. Schließlich wurde eine funktionale Analyse nach KEGG durchgeführt. Dabei wurde eine Überrepräsentation von RNA-Wegen in der RDI- und der PPI-Gruppe festgestellt, die in den meisten Fällen mit den Auswahlkriterien für RBPs übereinstimmte. Darüber hinaus haben wir auch eine Beteiligung solcher Interaktionspartner an Stoffwechsel- und Synthesewegen identifiziert. Weitere funktionale Auswirkungen für eine Untergruppe der nicht-essentiellen RBPs, die aus den 40 immunpräzipitierten RBPs extrahiert wurden, wurden mit einem Knockout (KO)-Proteom-Screen untersucht.

Die gesammelten Daten wurden mit manuell kuratierten *S. cerevisiae* Protein-Komplex-Annotationen kombiniert, um funktionsbasierte Netzwerke zu erstellen. So wurden Teilnetzwerke für die Funktionalitäten Spleißen, Export, Ribosom, Synthese, Stoffwechsel und Abbau erstellt. Aufgrund der Feststellung übereinstimmender RBP-Bindungsmuster wurde die netzwerkbasierte Zuordnung der RBP-Funktion vorgeschlagen. Das RNA-Splicing wurde als exemplarischer Prozess verwendet und weiterentwickelt, wodurch neuartige RBP-Assoziationen mit Spliceosom-Protein-Komplexen aufgedeckt wurden.

Insgesamt haben wir ein umfassendes RBP-Interaktom-Netzwerk bereitgestellt, das systematisch RDIs und PPIs identifiziert. Die gleichzeitigen Bindungsmuster innerhalb des Netzwerks ermöglichen die Vermutung funktionaler Verbindungen für die ausgewählten RBPs und deren Interaktionspartnern. Die Ressource ist im RBP-Interaktom-Netzwerk-Explorer (RINE) erfasst, der über eine Online-Interaktionsplattform unter <https://www.butterlab.org/RINE> zugänglich ist.

3.2.3 Statement of contribution

I was in charge of this project, including data collection and analysis. Hence, I planned and conducted most experiments and arranged the co-authors' contributions on specific experiments and data analysis. I prepared all the figures and wrote the first manuscript draft. Together, with Falk Butter, Emily Nischwitz, and Katja Luck we finalised the manuscript.

Supervision confirmation

Falk Butter

RNA-dependent interactome allows network-based assignment of RNA-binding protein function

Albert Fradera-Sola¹, Emily Nischwitz¹, Marie Elisabeth Bayer¹, Katja Luck² and Falk Butter^{1,*}

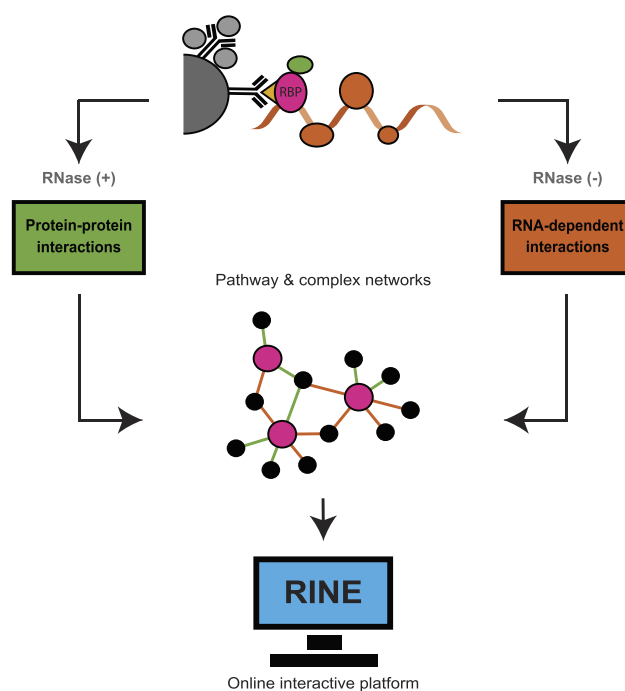
¹Quantitative Proteomics, Institute of Molecular Biology, D-55128 Mainz, Germany and ²Integrative Systems Biology, Institute of Molecular Biology, D-55128 Mainz, Germany

Received August 12, 2022; Revised March 16, 2023; Editorial Decision March 17, 2023; Accepted March 24, 2023

ABSTRACT

RNA-binding proteins (RBPs) form highly diverse and dynamic ribonucleoprotein complexes, whose functions determine the molecular fate of the bound RNA. In the model organism *Saccharomyces cerevisiae*, the number of proteins identified as RBPs has greatly increased over the last decade. However, the cellular function of most of these novel RBPs remains largely unexplored. We used mass spectrometry-based quantitative proteomics to systematically identify protein–protein interactions (PPIs) and RNA-dependent interactions (RDIs) to create a novel dataset for 40 RBPs that are associated with the mRNA life cycle. Domain, functional and pathway enrichment analyses revealed an over-representation of RNA functionalities among the enriched interactors. Using our extensive PPI and RDI networks, we revealed putative new members of RNA-associated pathways, and highlighted potential new roles for several RBPs. Our RBP interactome resource is available through an online interactive platform as a community tool to guide further in-depth functional studies and RBP network analysis (<https://www.butterlab.org/RINE>).

GRAPHICAL ABSTRACT



INTRODUCTION

Throughout their life cycle, mRNAs are bound by different RNA-binding proteins (RBPs) forming transient ribonucleoprotein (RNP) complexes. RNP complexes are critical to determine the downstream effects of the bound mRNAs (1). These downstream effects involve the initial mRNA processing, export from the nucleus, transport and localization within the cytoplasm and ultimate translation and degradation of the mRNA. mRNA processing mechanisms, including the initial steps of capping, splicing and polyadenylation, typically require RBP-mediated modifications of the mRNA. These modifications are later recognized by additional RBPs that trigger further coupled processes (2–4). Then, transmembrane RBPs play a critical role in

*To whom correspondence should be addressed. Tel: +49 6131 39 21570; Fax: +49 6131 39 21521; Email: f.butter@imb.de

facilitating the passage of mRNA from the nucleus into the cytoplasm (5,6). Once the mRNA has reached the cytoplasm, RNP complexes are again assembled in a sequential and contemporaneous manner to regulate mRNA cellular fate, such as localization, translation or degradation within a large interconnected network (7).

This web of connections is facilitated by interactions with a unique combination of RBPs, as either core or regulatory factors. Core factors are the central players in RNA processes and can be found to interact with a plethora of RNA species. For example, Pab1 is a critical core factor playing a central role in several steps of mRNA processing and metabolism (8,9). Regulatory factors, on the other hand, are more specific; they include, among others, the post-translational regulators interacting with specific sequences or structures of untranslated regions in mRNA (10). There is an interplay between core and regulatory factors, and the recruitment of the latter may result in the assembly of complexes that dictate exchange amid the core factors (11). Thus, the RBP combination of an mRNA determines its cellular fate.

Several large-scale approaches have been applied to discover RBPs in *Saccharomyces cerevisiae*. Previously, RBPs were identified by using protein arrays in which the capability of each arrayed protein to capture fluorescently labelled RNAs was measured (12,13). Additionally, mass spectrometry (MS) proteomics techniques were developed, including oligo(dT) capture (12) and *in vivo* RNA interactome capture (RIC), either with conventional cross-linking (14,15) or with photoactivatable ribonucleoside-enhanced cross-linking (16). Furthermore, MS-based techniques have been applied to identify RBPs by validation of a short RNA remnant fragment after cross-linking (17). Over the last decade, these studies have been consolidated into a census of 1273 proteins annotated as RBPs in *S. cerevisiae* (18). Within this census, there are a large number of proteins lacking canonical RNA-binding domains (RBDs), such as the eukaryotic RNA-recognition motif (RRM) or the heterogeneous nuclear RNP K homology (KH) domain. Of the seven studies included in the *S. cerevisiae* RBP census, the two largest contributors only reported 7% and 34% of proteins containing classical RBDs (15,16).

Among these proteins without canonical RBDs are a high number of metabolic enzymes. In recent years, there have been an increasing number of studies on the presence and evolutionary conservation of metabolic proteins having a secondary role as RBPs (19,20). In *S. cerevisiae*, 10% of the RBPs in the census are classified as metabolic enzymes (18). For instance, glyceraldehyde-3-phosphate dehydrogenase (21) and cytosolic aconitase (22,23) are well characterized to function as RBPs. This recurrent evidence showing metabolic enzymes acting as RBPs suggests an extensive enzyme activity regulation network acting through RNAs.

To better understand these RBPs, we need to accompany the RBP catalogue expansion with functional characterization of these proteins (24). Thus, it becomes paramount to correctly identify functional roles for the plethora of RBPs. One strategy used previously is to connect RBPs with a specific RNA sequence or structure to facilitate functional studies (25–27). Another strategy is to use the interconnection of RBPs binding to specific subsets of RNAs to dis-

entangle functionality (28). To implement this experimentally, RBPs can be immunoprecipitated in the presence or absence of RNA to describe concurrent interactions with other RBPs. This can be used to suggest the involvement of the bait RBP in functional pathways (29). Indeed, RNA-dependent protein interactors are more likely to be RBPs themselves, which has been used to predict RBPs from protein–protein interaction (PPI) networks (30). We reasoned that with sufficient data, this strategy can be extended to identify functional associations for previously uncharacterized RBPs. Thus, we immunoprecipitated 40 *S. cerevisiae* RBPs involved in different RNA pathways and identified their concurrent RNA-dependent and -independent interactors by quantitative proteomics. We further quantified proteome-wide protein expression level changes upon knockout of 13 of these RBPs. Integration of these data with pathway and protein complex annotations revealed new associations and functions of selected RBPs within core RNA maturation and regulation pathways, such as splicing.

MATERIALS AND METHODS

Yeast culture and lysis

Saccharomyces cerevisiae tandem affinity purification (TAP)-tagged strains (31) or knockout (KO) strains (32) (Horizon discovery) were grown for 2 days at 30°C on YPD agar plates. The resulting isolated colonies were inoculated on 15 ml of YPD medium and grown at 30°C and 180 rpm until saturation. Saturated cultures were spiked into 500 ml (RBP interactome screen) or 100 ml (KO screen) of YPD and grown (30°C and 180 rpm) until exponential growth (OD_{600} between 0.8 and 1.0 absorbance units), when cells were harvested at 3000 g for 5 min. Pelleted cells were suspended in 200 μ l of Buffer 1 [50 mM Tris pH 7.5, 150 mM NaCl, 5 mM MgCl₂ and freshly added 1 μ g/ml pepstatin and leupeptin, 1 mM phenylmethylsulphonyl fluoride (PMSF) and 0.5 mM dithiothreitol (DTT)] and transferred to a 2 ml screw lid tube containing 0.5 mm diameter zirconia/silica beads (Roth). Cells were lysed on a FasPrep-24 (MP Biomedicals) with two 30 s cycles at 6.5 m/s, allowing the samples to cool on ice in between. Cell lysates were topped up with 800 μ l of Buffer 2 (Buffer 1 + 0.2% IGEPAL), vortexed and transferred into a new tube, leaving the beads behind. Cell lysates were centrifuged at 15 g twice for 5 min at 4°C, and the supernatant was transferred into a clean tube after each cycle. Finally, the protein concentration was measured with a Bradford assay (Protein Assay Dye Reagent, Bio-Rad).

Immunoprecipitation

Protein G magnetic Dynabeads (30 mg/ml, Invitrogen) were separated with a magnetic rack and washed twice with 1 ml of Buffer 2 (see ‘Yeast culture and lysis’). Beads [20 μ l/immunoprecipitation (IP)] were coupled with 1 μ g/IP anti-TAP antibody (0.5 mg/ml, GenScript Biotech) in 500 μ l of Buffer 2 for 30 min on a rotating wheel at room temperature. Then, the beads were washed twice with 200 μ l of Buffer 2 and suspended in 100 μ l of Buffer 2. For each immunoprecipitation, 12 mg of protein lysate was combined with the 100 μ l of suspended beads and incubated for 3 h

on a rotating wheel at 4°C. Then, the beads were washed with 1 ml of Buffer 2 and split into two groups. One group was washed three times with 200 µl of Buffer 3 (Buffer 2 + 10% glycerol) containing 50 µg/IP RNase A from bovine pancreas (Sigma-Aldrich); the other was washed three times with 200 µl of Buffer 3 containing 0.5 µl/ml Ribolock RNase inhibitor (40 U/µl, Fisher Scientific). Finally, the beads were spun down and eluted with 30 µl of lithium dodecylsulphate (LDS) + 10 mM DTT.

MS sample preparation

Both RBP interactome screen immunoprecipitation elution products and KO screen protein lysates (100 µg in 30 µl LDS + 10 mM DTT) were heated for 10 min at 70°C. Proteins were then each separated on either a 4–12% (RBP interactome screen) or a 10% (KO screen) NOVEX gradient SDS gel (Thermo Scientific) for 8 min at 180 V in 1× MES buffer (Thermo Scientific). Proteins were fixed and stained with a Coomassie solution [0.25% Coomassie Blue G-250 (Biozym), 10% acetic acid, 43% ethanol]. The gel lane was cut into slices, minced and destained with a 50% ethanol/50 mM ammonium bicarbonate pH 8.0 solution. Proteins were reduced in 10 mM DTT for 1 h at 56°C and then alkylated with 50 mM iodoacetamide for 45 min at room temperature, in the dark. Proteins were digested with LysC (Wako Chemicals) overnight at 37°C. Peptides were extracted from the gel twice using a mixture of acetonitrile (30%) and 50 mM ammonium bicarbonate pH 8.0 solution, and three times with pure acetonitrile, which was subsequently evaporated in a concentrator (Eppendorf) and loaded on activated C18 material (Empore) StageTips as previously described (33).

MS data acquisition and analysis

Peptides were separated on a 25 cm self-packed column (New Objective) with a 75 µm inner diameter filled with ReproSil-Pur 120 C18-AQ (Dr. Maisch GmbH) with reverse-phase chromatography. The EASY-nLC 1000 (Thermo Fisher) was mounted on a Q-Exactive plus mass spectrometer (Thermo Fisher) and peptides were eluted from the column in an optimized 90 min (RBP interactome screen) or 4 h (KO screen) gradient from 2% to 40% MS-grade acetonitrile/0.1% formic acid solution at a flow rate of 200 nl/min. The mass spectrometer was used in data-dependent acquisition mode with one MS full scan and up to 10 MS/MS scans using HCD fragmentation. Raw MS data were searched using the Andromeda search engine (34) integrated into MaxQuant software suite 1.6.5.0 (35) using the S288C_Genome.Release.64-2-1_orf_trans_all.fasta protein sequences from the Saccharomyces Genome Database (SGD) (36). For the analysis, carbamidomethylation at cysteine was set as a fixed modification and methionine oxidation and protein *N*-acetylation were variable modifications. The match between runs option was activated.

Knockout library validation

For 18 of our 40 investigated RBPs, deletion strains are available in the *S. cerevisiae* KO collection (32). For the 18

available strains, we were able to validate the respective RBP knockout on the proteome level for 10 strains, visible as a strong down-regulation due to imputation of missing LFQ (label-free quantitation) values. In four cases, knockout validation was not possible because the target RBP was not detected in the wild type (WT). For the *cdc2-Δ* strain, Cdc2 expression levels in the KO strain were equal to those in the WT strain. We thus decided to check all strains by colony polymerase chain reaction (PCR) targeting the respective open reading frame (ORF) and the presence of an incorporated kanamycin resistance marker, which should be present in all KO clones.

KO and WT strains were streaked on YPD agar plates and grown for 3 days at 30°C. A single colony was re-suspended in 10 µl of fresh 0.02 N NaOH, and incubated in PCR tubes at 99°C for 10 min. The supernatant was transferred to a fresh tube, and chilled on ice for 10 min. To amplify the samples, OneTaq (NEB, M0480S) was used according to the manufacturer's instructions, containing 2 µl of yeast lysis supernatant in a 25 µl reaction. Horizon discovery 'YKO Primers from SGDP' were used. A_confirmation_primer and B_confirmation_primer were used to detect the WT allele, and A_confirmation_primer and kanB were used to detect the KO allele. The annealing temperature was set at 50°C and used with a 90 s elongation cycle. Samples were separated on a 1% agarose gel and imaged with Gel Doc™ XR+ (Bio-Rad) with a 1 s exposure.

This validated 13 KO strains—the 10 previously validated with proteome evidence and another three previously not detected by MS. The two strains without validation (*cdc2-Δ* and *mssl1-Δ*) were excluded from further downstream analysis.

Bioinformatics analysis

For protein quantification, contaminants, reverse database hits, protein groups only identified by site and protein groups with <2 peptides (at least one of them classified as unique) were removed by filtering from the MaxQuant proteinGroups.txt file. Missing values were imputed by shifting a beta distribution, obtained from the LFQ intensity values, to the limit of quantitation. Further analysis and graphical representation were performed on an R framework (37) incorporating ggplot2 (38), for visualization, among other packages. For the RBP interactome screen, the protein enrichment threshold was set to a *P*-value <0.05 (Welch *t*-test) and fold change >2, *c* = 0.05. Enriched proteins were overlapped with all interactors with physical evidence reported at The Biological General Repository for Interaction Datasets (BioGRID) (39) and with proteins in the RBP census (18). For the KO screen, the protein enrichment threshold was set to a *P*-value <0.05 (Welch *t*-test) and to abs(fold change) >2, *c* = 0.05.

For the protein signature enrichment analysis, domains were queried with InterProScan version 5.50-84.0 (40), and hits in the Pfam (41) and SUPERFAMILY (42) databases were selected for downstream analysis. Signatures for a particular group of enriched proteins were tested for over-representation (*P*-value <0.01; Fisher's exact test) against all signatures found in the background (defined as all quantified proteins in the comparison, enriched or not).

Signatures found to be over-represented in at least two bait RBPs were selected for graphical representation.

For the molecular function enrichment analysis, terms were queried in the Gene Ontology (GO) database (43) with the ClusterProfiler R package (44). Terms for a particular group of enriched proteins were tested for over-representation {adjusted P -value [false discovery rate (FDR)] <0.05 ; Fisher's exact test} against all terms found in the background (defined as all quantified proteins in the comparison, enriched or not). The top five most found terms per group, which were over-represented in at least two bait RBPs, were selected for graphical representation.

For the pathway enrichment analysis, terms were queried in the Kyoto Encyclopedia of Genes and Genomes (KEGG) (45) and Reactome (46) databases with the ClusterProfiler R package (44). Terms for a particular group of enriched proteins were tested for over-representation [adjusted P -value (FDR) <0.05 ; Fisher's exact test] against all terms found in the background, defined as all other quantified proteins in the comparison.

For the protein complex analysis, enriched proteins were overlapped with the manually curated heteromeric protein complexes included in the CYC2008 data (47). A ratio for each complex was calculated by dividing the number of proteins overlapping with a particular complex by the total number of proteins in that complex. For the RBP interactome screen data, protein complexes with a ratio >0.5 were selected for graphical representation. Meanwhile, for the KO screen data, the threshold was set at a ratio ≥ 0.5 .

Finally, protein networks were generated with in-house scripts based on an R framework incorporating igraph (48), with the Fruchterman–Reingold force-directed layout algorithm implementation, among other packages. All networks were drawn with the spoke model. For the PPI and RNA-dependent interaction (RDI) global networks, bait RBPs with an associated KEGG term and their prey were selected as nodes. For the functional subnetworks, bait RBPs and their prey were included as nodes when associated with a particular functionality via KEGG analysis, even when associated with multiple terms. Preys were then coloured in grey tones when reported to BioGRID (to any of its interacting baits) and blue tones when not, and with darker tones when reported at the RBP census and lighter tones when not. For the complex subnetworks, bait RBPs and their prey were selected as nodes when interacting or being part of a particular complex. The network explorer interactive platform was developed with the Shiny R package (49).

RESULTS AND DISCUSSION

Quantitative interactomics screen identifies protein–protein and RNA-dependent interactions

We selected RBPs from pathways that span the RNA life cycle, including less characterized RBPs from recent RIC studies (14–16). These RBPs are involved in seven major RNA-associated processes: (i) capping; (ii) splicing; (iii) cleavage; (iv) polyadenylation; (v) nuclear export; (vi) transport and localization; and (vii) degradation (Figure 1A). To select the individual RBPs for the interactome screen, we queried the RBP census (18) for proteins annotated with

these roles in the KEGG and Reactome databases (Supplementary Figure S1). Notably, we included 40 RBPs covering various stages of the mRNA life cycle either with specialized roles in particular pathways or with broader functional descriptions (highlighted with an asterisk in Figure 1A). Of these 40 RBPs, two (Spt5 and Sto1) are involved in capping, 14 in splicing, three (Cft1, Mpe1 and Rna14) in cleavage, Pab1 in polyadenylation and Puf3 in transport/localization. Furthermore, two RBPs (Ndc1 and Mex1) were selected for nuclear export and 17 RBPs are associated with RNA degradation.

We performed a quantitative label-free proteomics screen with the 40 chosen TAP-tagged RBPs (31) using two different conditions and a WT (Figure 1B), which allowed us to differentiate between protein–protein associations/interactions (PPIs) and RNA-dependent associations/interactions (RDIs). To identify the PPIs, we compared immunoprecipitated tagged RBP against WT lysate, both treated with RNase A to digest the RNA, similar to previous large-scale yeast PPI screens (50,51). We also included a second condition where we immunoprecipitated the tagged RBP, but omitted the RNase A treatment, which reveals the RDIs that are only observable in the presence of RNA. Each condition comprised multiple replicate IP experiments that were prepared in parallel and measured on the mass spectrometer as a set applying LFQ.

This study design allowed for quality control benchmarking within each IP set using the bait RBP and RNase A treatment. In the case of RBP-IP compared with the WT lysate, the tagged RBP was expected to be enriched (P -value <0.05 and fold change >2 , $c = 0.05$) (Figure 2A). Indeed, 38 of 40 bait RBPs showed strong enrichment, between 3.3- and 14.1-fold (Supplementary Figure S2). The remaining two tagged RBPs, Ndc1 and Spt5, also showed enrichment of 2.4- and 2.1-fold, respectively, despite slightly less statistical significance of P -value = 0.07 and P -value = 0.10. Additionally, when comparing the RBP IPs with and without RNase treatment, the tagged RBP is expected to be equally abundant (Figure 2B). Again, this was the case for almost all experiments (39 of 40), with only Sub2 showing a slight offset (Supplementary Figure S3). Similarly, we clearly see that RNase A is found in the non-enriched background cloud of proteins, when comparing the RBP IPs with RNase treatment [IP RNase (+) versus WT]. Meanwhile, we observed the RNase A enriched, with a negative fold change, when comparing the non-treated with the treated RBP IPs [IP RNase (–) versus IP RNase (+)] (Figure 2A, B; Supplementary Figures S2 and S3).

We obtained valuable information for the PPIs (Figure 2C; Supplementary Table S1) and RDIs (Figure 2D; Supplementary Table S2) for the 40 chosen RBPs. For the PPI group, the number of enriched proteins ranged from 4 (Dbp2) to 112 (Mpe1) (Figure 2C), while for the RDI group, the number of enriched proteins ranged from 5 (Upf3) to 143 (Nam7) (Figure 2D). We then used BioGRID, a database of established protein interactions, to check how many of the identified interactions had been previously reported with physical experimental evidence (light grey, left side, Figure 2C and D). The ratio of reported interactions was dependent on the bait RBP, ranging from 0 (for Dbp2, among others) to 80% (for Dhh1) for PPIs and from 0 (for

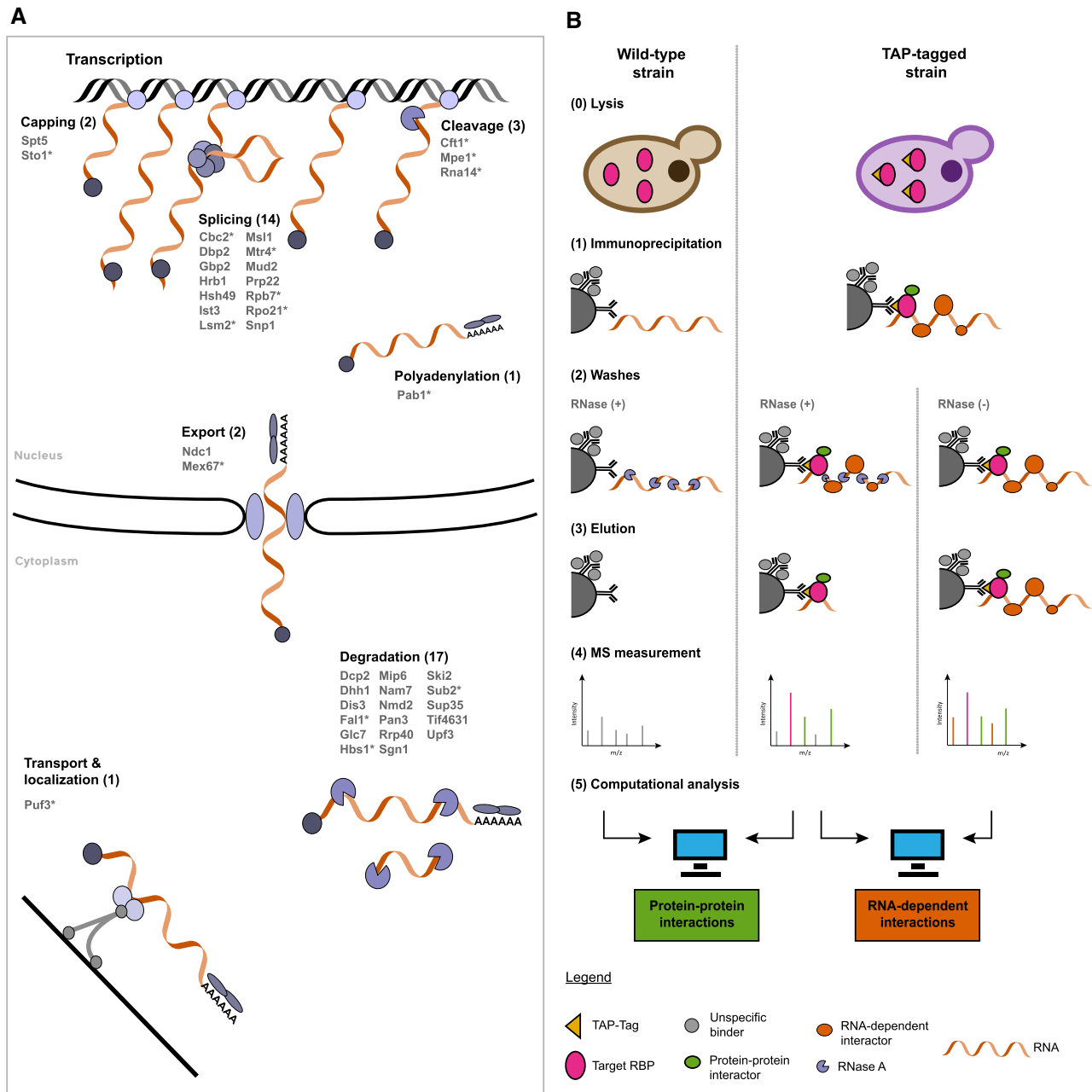


Figure 1. RBP interactome screen. (A) The 40 selected bait RBPs are listed with a schematic drawing of their RNA biological processes. Proteins highlighted with an asterisk are associated with multiple processes. (B) Schematic representation of the experimental design to screen for protein–protein interactors and RNA-dependent interactors in parallel.

Dbp2, among others) to 82% (for Dhh1) for RDIs. As expected, the percentage of previously identified interactions in the BioGRID database was overall higher for PPIs than for RDIs. For 11 RBPs, our PPIs included at least half of the previously described interactions, while for our RDIs, this was the case for only two bait RBPs (Supplementary Figure S4A, B). Additionally, for 26 of our investigated RBPs, BioGRID classifies more than half of the reported interactions as ‘Affinity Capture-MS’, confirming the experimental results of our approach (Supplementary Figure S4C). Irrespective of PPIs or RDIs, we found high overlap with

RBPs included in the published RBP census (18) (black, right side, Figure 2C, D). However, the fraction of RBP-annotated proteins was higher within RDIs compared with PPIs. While the RDI partners of 34 bait RBPs consisted of >70% of RBPs, only 14 bait RBPs showed this high fraction for their PPI partners (Supplementary Figure S4D). Finally, when we checked the overlap of the interactors among the PPI and RDI groups of each bait RBP, we observed that 0 (for Dbp1, among others) to only 22% (for Sto1) are identical (dark grey, Figure 2E), clearly showing that they are two specific subsets of interactors.

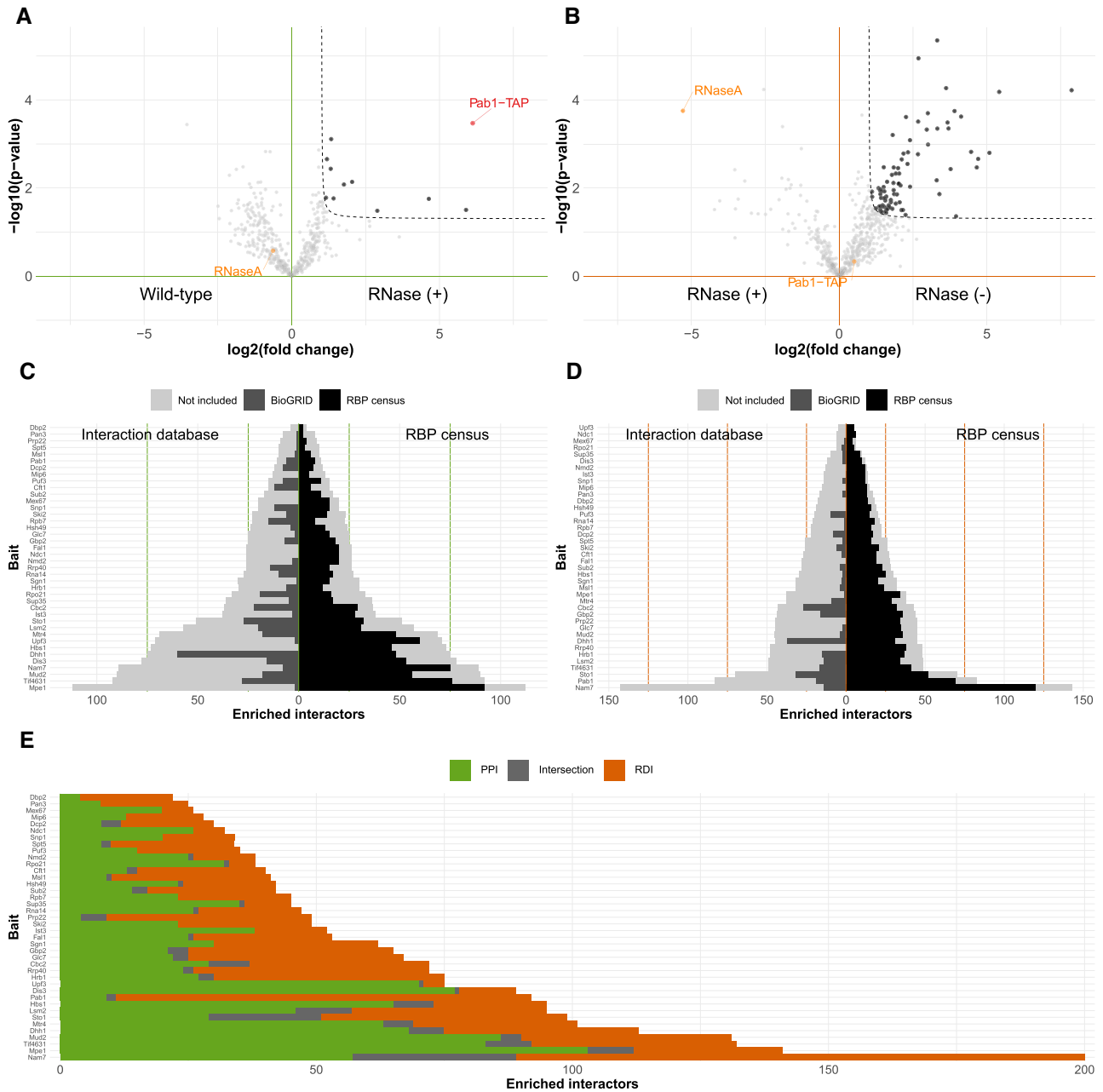


Figure 2. RBP interactome screen reveals different PPIs and overlapping RDIs among bait RBPs. (A) Volcano plot of PPIs for Pab1 comparing enriched proteins of the Pab1-TAP IPs digested with RNase A ($n = 4$) or WT ($n = 4$) determined by label-free quantitative proteomics. The enrichment threshold (dotted line) is set to P -value < 0.05 (Welch t -test) and fold change > 2 , $c = 0.05$. Each dot represents a protein; enriched proteins are shown in black. Pab1-TAP (red) and RNaseA (orange) are indicated. (B) Volcano plot of RDIs for Pab1 comparing enriched proteins of Pab1-TAP IPs ($n = 4$), with or without RNase A digestion, determined by label-free quantitative proteomics. The enrichment threshold (dotted line) is set to P -value < 0.05 (Welch t -test) and fold change > 2 , $c = 0.05$. Each dot represents a protein; enriched proteins are shown in black. Pab1-TAP and RNaseA are not enriched (orange). (C and D) Bar plot of PPIs (C) and RDIs (D) for the 40 bait RBPs. Each bar represents the number of enriched proteins [P -value < 0.05 (Welch t -test) and fold change > 2 , $c = 0.05$]. Each bar is mirrored to show the protein's overlap with reported interactors at the BioGRID database (left side, dark grey) and with the RBP census (right side, black). Proteins not contained in either are coloured in light grey. (E) Bar plot depicting all enriched interactors of the 40 bait RBPs: unique PPI (green), unique RDI (orange) and shared interactors (grey).

Overall, our label-free quantitative RBP interactome screen resulted in two distinct groups of enriched interactors among the PPI and RDI datasets. In the case of RDIs, the majority of protein interactors were included in the RBP census, outlining that our approach is able to uncover hitherto unknown RDIs among a large set of RBPs. With this, we provide complementary information to the previous large-scale screens in yeast that were designed to only report RNA-independent PPIs.

RNA-related functionalities are over-represented among enriched interactors

We wanted to investigate whether RNA functionalities were over-represented and shared among our enriched interactors, from the structural protein domain to the functional pathway level. Thus, we queried for protein signatures among the enriched interactors with InterProScan. For this analysis, we noted that the descriptors ‘RNA recognition motif domain’ and ‘RNA-binding’ (Figure 3A) were significantly over-represented (P -value <0.01) among multiple bait RBPs for both PPIs and RDIs in the structural domain databases Pfam and SUPERFAMILY. Within the PPIs, there were three baits with enrichment of these terms (green, Figure 3A; Supplementary Table S3). We then applied the same protein signature analysis for the enriched RDIs. RNA binding-related domains were most prevalent among the over-represented protein domains (orange, Figure 3A; Supplementary Table S4). For instance, ‘RRM’ (Pfam) and ‘RNA-binding domain’ (SUPERFAMILY) were over-represented among the interactors of 17 and 18 of our 40 bait RBPs, respectively. Overall, this shows a general trend of the RDIs having a larger amount of canonical RBDs.

Further interrogation of the enriched PPI set using GO revealed an over-representation [adjusted (FDR) P -value <0.05] of molecular function GO terms such as ‘nucleic acid binding’ among the interactors of nine bait RBPs and ‘RNA-binding’ among the interactors of six bait RBPs (Figure 3B; Supplementary Table S5). Over-represented RNA-related GO molecular function terms were also identified for the enriched RDI set and were more predominant than in the PPI set (Figure 3B; Supplementary Table S6). In particular, the term ‘mRNA binding’ was over-represented among the enriched RDIs for 26 of our 40 bait RBPs. Despite the domain and GO molecular function analysis revealing an enrichment for RNA binding functionalities, especially for RDI partners, the number of canonical RBDs among the bait RBPs’ interactors is low (Supplementary Tables S3 and S4) albeit still significant for the RRM domain (Figure 3A; Supplementary Figure S5). This highlights that among our enriched interactors, RBPs lacking canonical RBDs might be abundant and thus may have less studied functions in the context of RNA biology.

To investigate shared functionalities among our interactors, we queried the KEGG for over-represented pathways among interactors in the PPI and RDI datasets [adjusted (FDR) P -value <0.05 ; Supplementary Tables S7 and S8, respectively]. As expected, for both PPIs and RDIs, we obtained over-represented KEGG terms associated with known bait RBP functionality in several cases (Figure 3C).

In particular, within the 14 selected splicing-associated RBPs, the KEGG term ‘spliceosome’ was over-represented among PPIs of the bait RBPs Snp1, Ist3 and Hsh49, as well as for the RDIs of the bait RBP Msl1. For Cbc2 and Lsm2, the ‘spliceosome’ term was over-represented for both PPIs and RDIs. The degradation-associated RBPs Rrp40, Dhh1, Dis3, Sup35 and Hbs1 had RNA degradation-associated KEGG terms over-represented among their PPIs, while Sgn1 and Tif4631 showed this among their RDIs. We also enriched interactors related to ‘ribosome’ and ‘ribosome biogenesis in eukaryotes’ in four RDI (Cbc2, Mtr4, Mud2 and Tif4631) and two PPI (Gbp2 and Hsh49) datasets. Additionally, while not among our selection criteria for the bait RBPs, we observed a strong over-representation of metabolic and synthesis pathways among the interactors for 7 and 13 baits of the PPI and RDI groups, respectively. There were 525 metabolism-related proteins within the combined PPI and RDI datasets, which included 70 of the 154 metabolic RBPs described in the RBP census. Previously, metabolic proteins have been characterized to have unusual RNA binding function, which coincides with the lack of enrichment of canonical RBDs across all RBPs (18). Through our unbiased, global analysis, our dataset adds more evidence to the growing field of metabolic enzymes with RBP functionalities (18,20).

Overall, despite a high number of interactors without canonical RBDs, we confirm that interactors of both PPI and RDI groups are involved in RNA biology through association with structural domains, molecular functions and biological pathways. This is in agreement with the SONAR dataset, where they already used this observation on a smaller set of baits for the prediction of hitherto unknown RBPs (30).

RBP knockouts reveal possible processes regulated by these RBPs

RBPs that interact with a certain set of mRNAs have sometimes been associated with the post-transcriptional regulation of genes belonging to a specific biological process or protein complex (10,52). To further investigate the downstream biological processes that are likely to be regulated by the action of our bait RBPs, we aimed to identify proteins that are differentially expressed at the protein level upon knockout of our selected RBPs. For 18 of our 40 investigated RBPs, deletion strains were available in the *S. cerevisiae* KO collection (32). Of the 22 unavailable KO strains, 21 correspond to essential genes and one is not included in the library. Further experimental validation of the 18 available strains resulted in the confirmation of RBP knockout in 10 strains at the proteomic level, and an additional 3 strains at the genomic level (Supplementary Figures S6 and S7A, B; see also the Materials and Methods). Thus, 13 RBP strains were utilized for further experimental investigations; these included five RBPs involved in mRNA nuclear processing, one in RNA transport and localization, and seven in degradation (Supplementary Figure S1A).

We measured the proteomes of WT and individual KO clones by MS and performed label-free quantification with multiple replicates (Figure 4A). Per knockout, we quantified between 2854 and 3184 proteins (Supplementary Table S9),

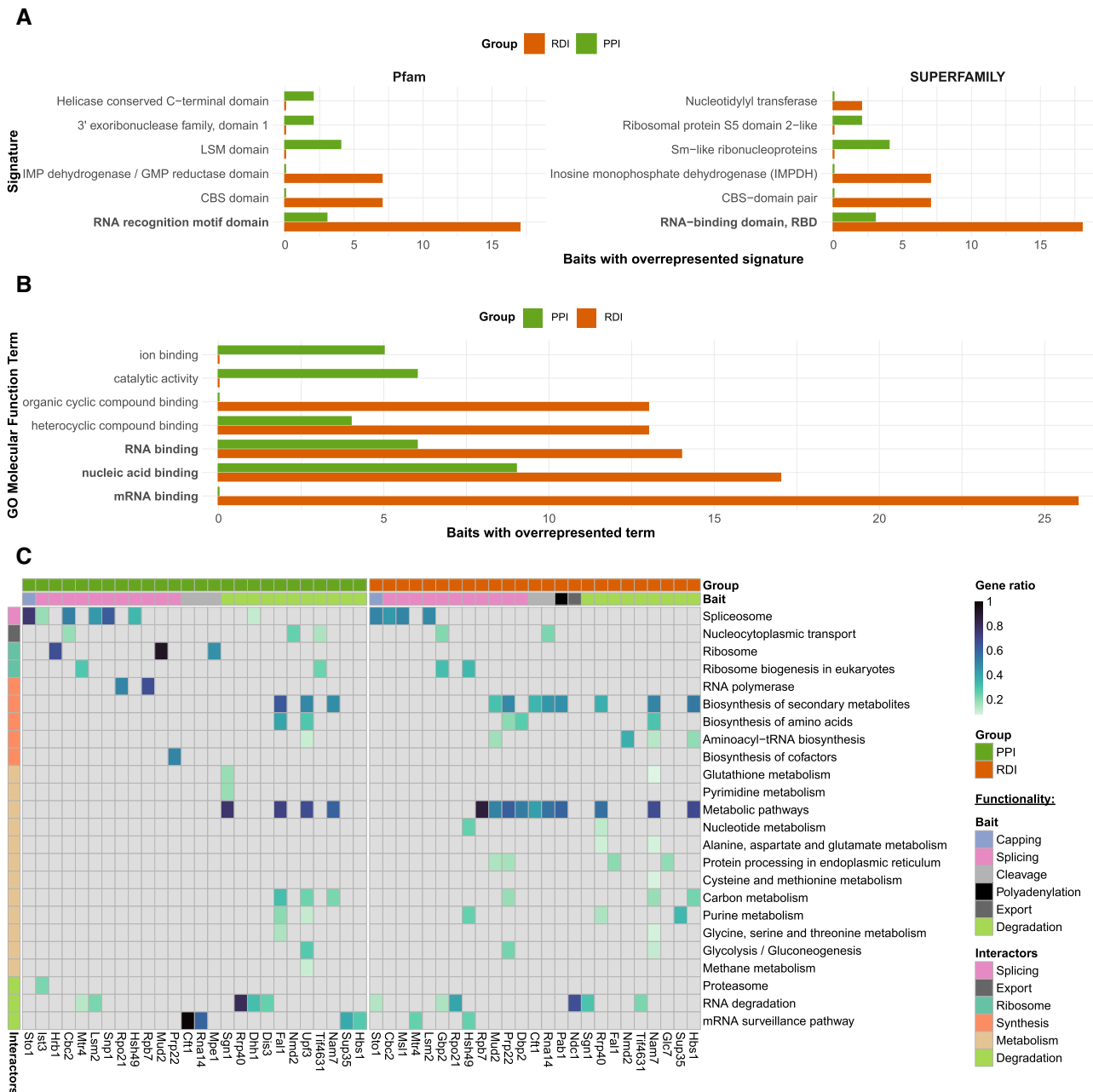


Figure 3. Computational analysis links the enriched interactors with RNA-related functionalities. **(A)** Bar plot of PPI (green) and RDI (orange) protein signatures. Each bar represents the number of bait RBPs with the over-represented signature (P -value <0.01 ; Fisher's exact test) found for at least two different bait RBPs. RNA binding-related signatures are in bold. **(B)** Bar plot of the PPI (green) and RDI (orange) GO molecular function terms. Each bar represents the number of bait RBPs with the over-represented term [adjusted (FDR) P -value <0.05 ; Fisher's exact test] for at least two bait RBPs. The top five terms per group are shown. RNA binding-related signatures are in bold. **(C)** Heat map of the PPI (green) and RDI (orange) KEGG analysis. Each row contains an over-represented KEGG term [adjusted (FDR) P -value <0.05 ; Fisher's exact test], with a blue colour gradient representing the gene ratio. The second horizontal bar represents the bait RBP functional selection criterion. The vertical bar represents the global function of the KEGG terms associated with the interactors.

approximately two-thirds of the expressed yeast proteome (53). To determine significant protein expression changes, we compared the WT with the KOs, setting a threshold at P -value <0.05 and $\text{abs}(\text{fold change}) >2$, $c = 0.05$ (Figure 4B). For the 13 RBP KOs, significant protein expression changes ranged from 7 to 230 proteins, representing on average a higher number of differentially expressed proteins than observed for a genome-wide KO screen performed in

Schizosaccharomyces pombe (54) and being on a par with expression changes observed for their *S. pombe* homologues (Figure 4C; Supplementary Figures S6 and S8; Supplementary Table S10). This shows that the knockout of the selected RBPs led on average to more profound expression changes than knockout of other genes. As expected, we observed that differentially expressed proteins overlapped little with prey identified in the PPI and RDI datasets of each

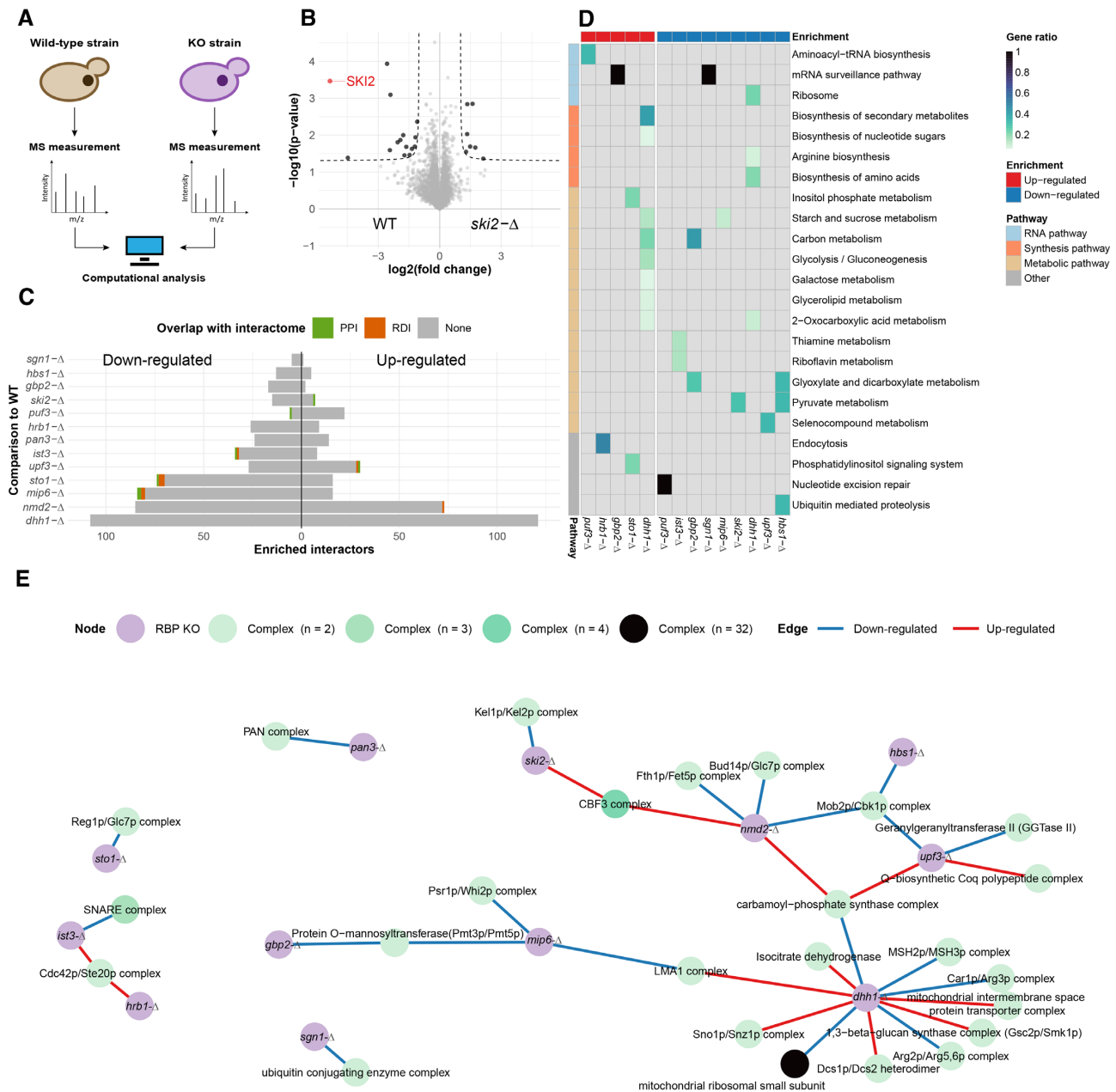
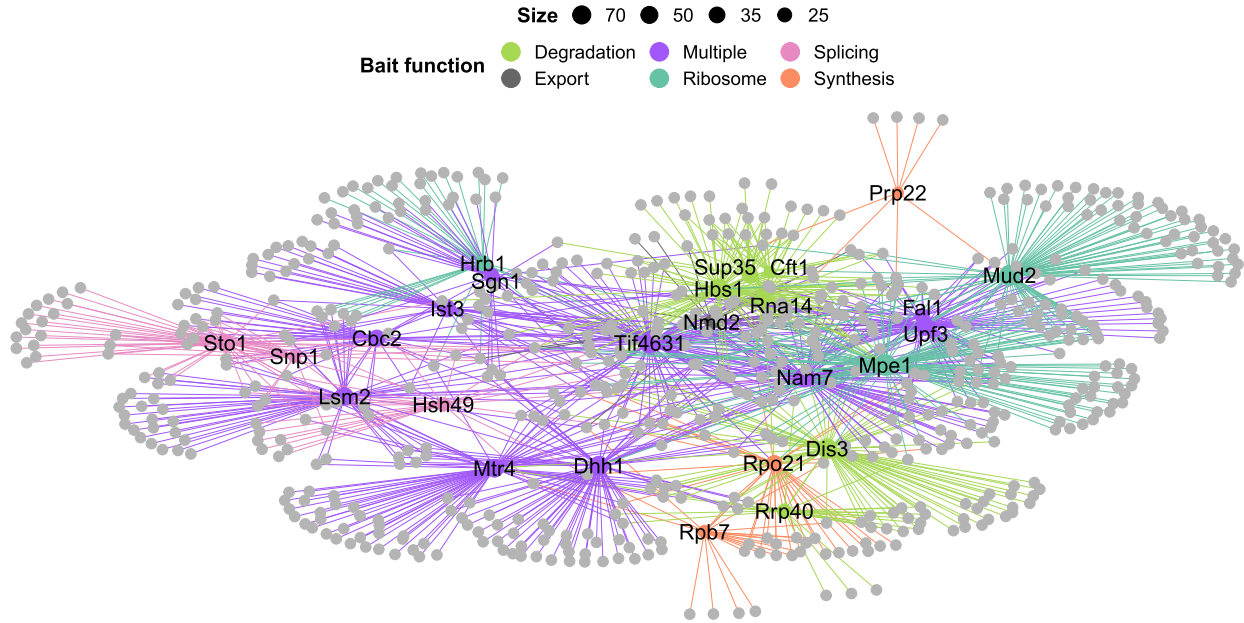


Figure 4. Protein expression changes among the 13 RBP KO strains. (A) Schematic representation of the KO screen experimental design. (B) Volcano plot of *ski2-Δ* comparing its proteome with that of the WT by label-free quantitative proteomics ($n = 4$). The enrichment threshold (dotted line) is set to P -value < 0.05 (Welch t -test) and $\text{abs}(\text{fold change}) > 2$, $c = 0.05$. Each dot represents a protein; enriched proteins are shown in black with *Ski2* highlighted in red. (C) Bar plot of the altered proteins in the 13 RBP KO strains. Each bar shows the number of altered proteins with highlighted overlap for PPI (green) and RDI (orange) with the RBP interactome screen. (D) Heat map of the KEGG analysis for the 13 RBP KO strains. Each row contains an over-represented KEGG term [adjusted (FDR) P -value < 0.05 ; Fisher's exact test] with a blue colour gradient for the gene ratio. The horizontal bar delineates up- and down-regulated proteins. (E) Network of the protein complexes with at least half of their subunits included among the differentially expressed proteins in the RBP KO strains. Nodes are RBP KO (purple) and protein complexes (blue gradient). Edges are highlighted for up-regulated (red) or down-regulated (blue) proteins.

respective bait RBP; ranging from 0 (for *gpb2-Δ*, among others) to 4.4% (for *sto1-Δ*) (Figure 4C). This supports our initial hypothesis that the differentially expressed proteins in the KO strains point to downstream processes regulated by the RBPs and are thus different from the RBP-associated proteins. However, differential expression of some proteins can also be the result of functional compensation mechanisms.

To further characterize the biological pathways that are affected and likely to be regulated by the knocked out RBPs, we tested for KEGG pathway enrichment among the up- and down-regulated proteins separately in each strain (Figure 4D; Supplementary Tables S11 and S12). Only three pathways related to RNA were over-represented in three KO strains, 'aminoacyl-tRNA biosynthesis', 'mRNA surveillance pathway' and 'ribosome' (Figure 4D).

A



B

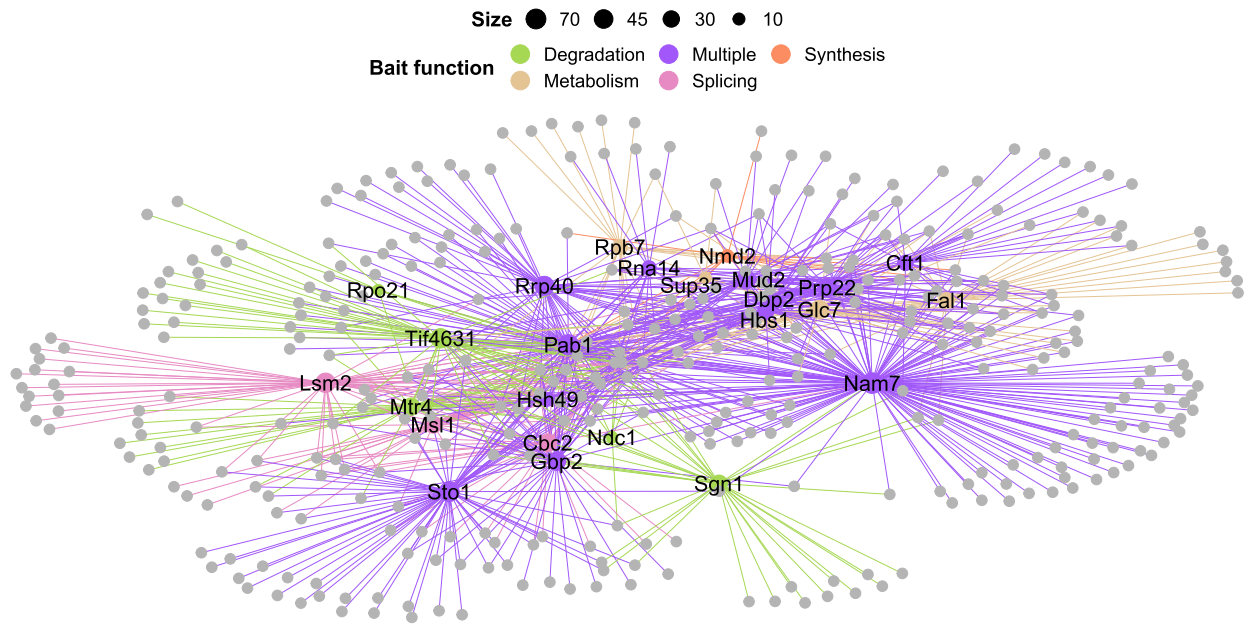


Figure 5. PPI and RDI networks link novel interactors to RNA-related functionalities. (A and B) PPI network (A) and RDI network (B) with all bait RBPs labelled. Bait RBP nodes and edges are coloured according to single (colour indicated in the key) or multiple (purple) higher order KEGG-associated functions. The nodes of the RBP prey are light grey and node sizes are determined by the number of interactors. The networks are drawn with the spoke model.

Interestingly, amino acid and nucleic acid synthesis pathways as well as various metabolic pathways were enriched among proteins that were down-regulated in KOs of RBPs related to splicing (*ist3-Δ* and *gpb2-Δ*) and to degradation (*sgn1-Δ*, *mip6-Δ*, *ski2-Δ*, *dhh1-Δ*, *upf3-Δ* and *hbs1-Δ*). However, these do not overlap with the over-represented metabolic KEGG pathways from the RBP interactome screen.

To further interrogate cellular processes that are likely to be regulated by the selected RBPs, we identified known yeast protein complexes for which at least half of all subunits were either up- or down-regulated in the individual KO strains (Figure 4E). We primarily obtained dimeric complexes (27 of 31) across the 13 selected RBPs. These complexes covered a wide range of functionalities. Some were expected, such as the down-regulation of

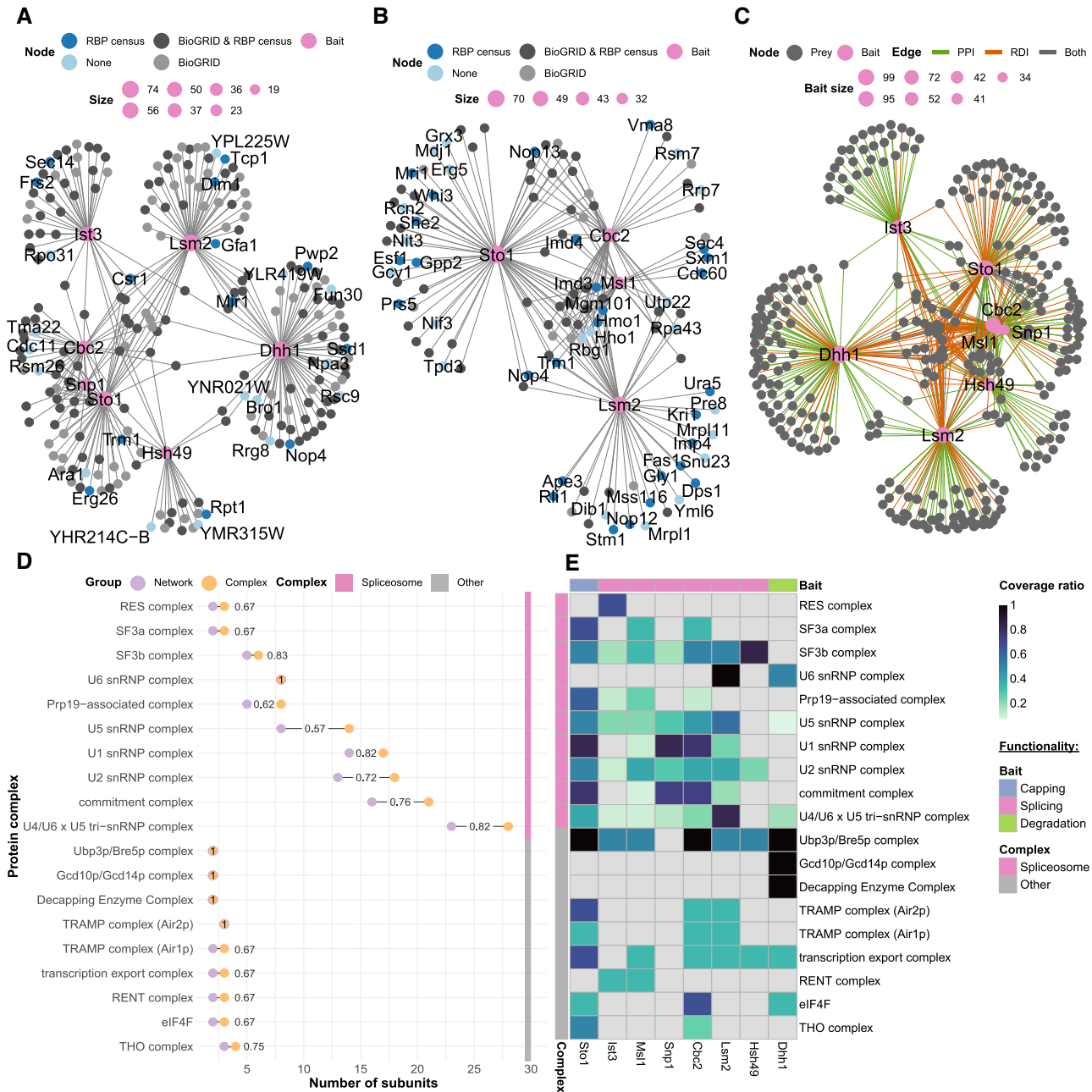


Figure 6. Splicing subnetworks link bait RBPs with protein complexes. (A and B) PPI splicing (A) and RDI splicing (B) subnetworks, with bait RBPs possessing interactors, with KEGG-associated splicing pathways highlighted in pink. Interactors reported at BioGRID (light grey), at the RBP census (dark blue) or at both (dark grey) are indicated. Interactors not found in BioGRID or the RBP census are indicated in light blue. The node size of the bait RBPs is determined by the number of interactors. The networks are drawn with the spoke model. (C) Combined (PPI and RDI) network, with bait RBPs having interactors with KEGG-associated splicing pathways in pink. Edges are highlighted for unique PPI (green), unique RDI (orange) or shared interactors (grey). The network is drawn with the spoke model. (D) Protein complex Cleveland dot plot. Protein complexes with a coverage ratio >0.5 are represented. The number of proteins in the network (purple) and in the complex (orange) are shown as dots. (E) Heat map of the protein complexes. Each row names a protein complex, with a blue colour gradient for the coverage ratio. The horizontal bar indicates the bait RBP functional selection criterion.

the PAN complex that is directly related to the knockout of one of its subunits, Pan3. Others echoed with the over-represented KEGG pathways. For instance, we obtained several metabolism-related complexes that were up-regulated in *dhh1*-Δ. Similarly, *mip6*-Δ and *gpb2*-Δ, both having KEGG metabolic pathways down-regulated, had a subunit of the Pmt3p/Pmt5p complex down-regulated. Ad-

ditionally, we obtained further insights into RBPs when examining the up- and down-regulated complex members. In *dhh1*-Δ we found a down-regulation of mismatch repair proteins. This is in line with the involvement of Dhh1 in DNA repair (55). Dhh1 has also been shown to be critical to G₁/S phase cell cycle progression, and its deletion causes ionizing radiation sensitivity (56). There are

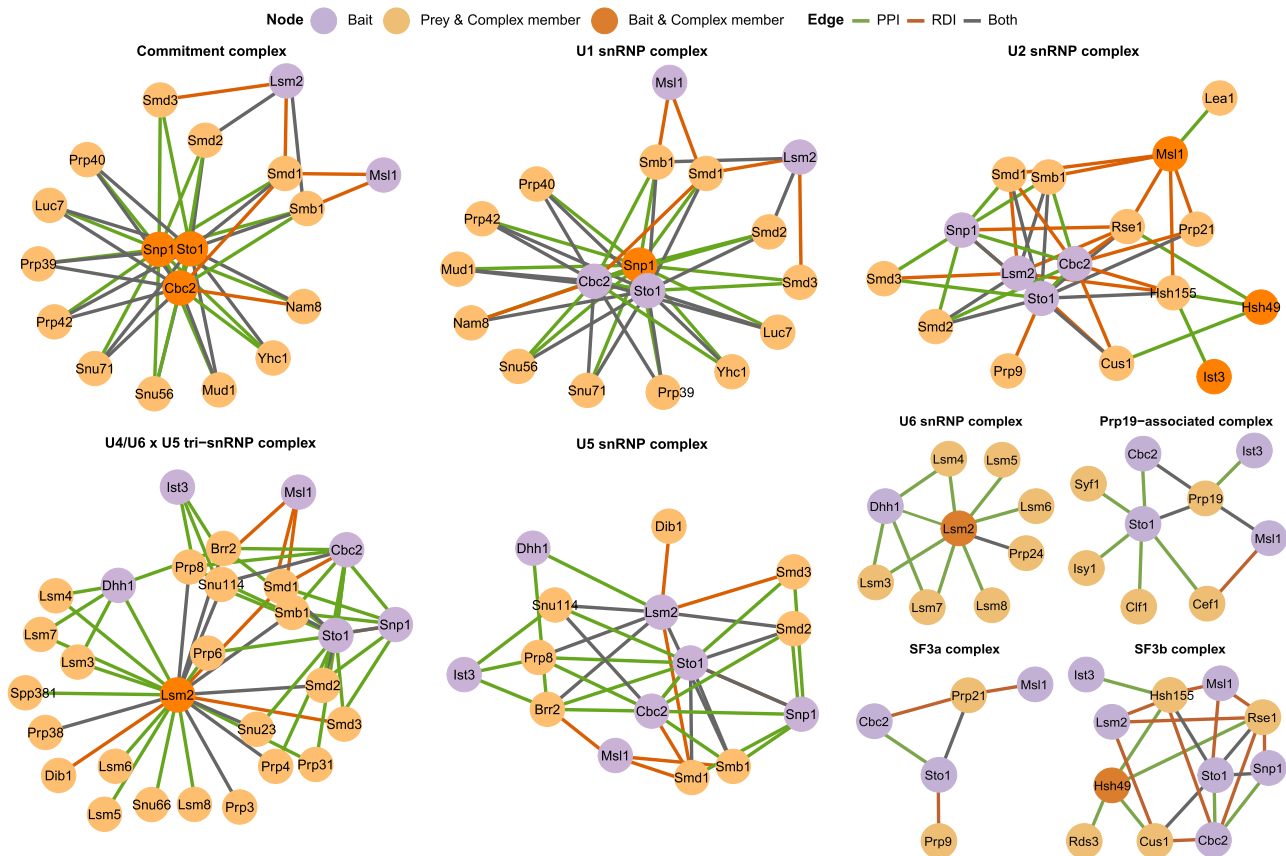


Figure 7. Spliceosome protein complex subnetworks reveal novel bait RBP functionalities. Spliceosome complexes shared by more than one bait RBP and with a coverage ratio >0.5 are shown. Bait RBPs being part of (dark orange) or associated with (purple) a protein complex are shown. Prey being part of a protein complex (light orange) are shown. Edges are coloured for PPI (green), RDI (orange) or both (grey). The networks are drawn with the spoke model.

emerging studies that have linked RBPs with DNA repair processes (57,58). Taken together, this KO dataset could lead to the further association of RBPs with metabolism and DNA repair.

These results collectively point to interesting hypotheses for possible downstream processes regulated by the selected RBPs and can guide future experimental investigations.

Network analysis identifies putative new members of RNA pathways

To connect information from the individual experiments and visualize shared interactors among the RBPs, we built an extensive interaction network for the PPI (Figure 5A; Supplementary Table S13) and RDI (Figure 5B; Supplementary Table S14) datasets. Within each network, the bait RBPs and their interactors were included when enriched for a KEGG term (Figure 3C). These KEGG terms were grouped as degradation, export, metabolism, ribosome, splicing, synthesis or multiple. Each of these individual categories was used to create specific subnetworks. To gain further insights into which complexes are captured within our subnetworks, we annotated our proteins with the manually curated heteromeric protein complexes included in the CYC2008 dataset (47). We calculated for each complex a coverage ratio by dividing the number of sub-

units included in the network by the total number of complex members, and those with a ratio >0.5 were included in the downstream analysis. Within all splicing, export, ribosome, synthesis, metabolism and degradation subcomplexes, 56 unique complexes were identified. These combined networks as well as individual subnetworks are available online for each enriched biological process within the RBP interactome network explorer (RINE) at <https://www.butterlab.org/RINE>.

We wanted to check the presence of expected and unexpected complexes as well as cross-talk between these complexes in an exemplary RNA process. Thus, we further examined the PPI and RDI subnetworks of the baits enriched for splicing-related KEGG terms, which have been widely studied in *S. cerevisiae* (59). Within the interactions, there were seven baits included in the PPI subnetwork (Ist3, Lsm2, Cbc2, Dhh1, Snp1, Sto1 and Hsh49) (Figure 6A; Supplementary Table S15), and there were four bait RBPs included in the RDI subnetwork (Ist3, Cbc2, Msl1 and Lsm2) (Figure 6B; Supplementary Table S16). We had 15 and 20 previously unreported interactions in the PPIs and RDIs, respectively, when compared with BioGRID and RBP census data (Figure 6A, B). Among these PPIs, there were three uncharacterized interactors, namely YHR214C-B and YMR315W with bait RBP Hsh49, and YPL225W with bait RBP Lsm2. We additionally integrated both the

PPI and RDI interactomes into one combined subnetwork (Figure 6C; Supplementary Table S17).

Using the protein complexes from the CYC2008 dataset, we established complexes present at a ratio >0.5 within the combined RDI and PPI subnetwork (Figure 6D; Supplementary Table S18). There were 19 complexes within the PPIs and RDIs with over-represented splicing-related KEGG pathways that surpassed this threshold. These complexes contained the anticipated spliceosome components as well as other complexes that are involved in the mRNA life cycle. The complexes that were not spliceosome components function in capping, degradation, nuclear export and translation (Figure 6D). For example, the transcription export (TREX) complex, which contains the THO complex, were both found among the interactors. These complexes are critical for the nuclear export of mRNA (60). There were six baits with interactors that are components of the TREX complex (Figure 6E). Of these, Msl1, Snp1 and Hsh49 are considered splicing-associated proteins, while Dhh1 is considered to be degradation associated. This shows the interconnectedness of mRNA splicing, export and degradation. An unexpected complex was the RENT complex, which is responsible for rDNA silencing in *S. cerevisiae*. The bait RBPs Ist3 and Msl1, which are both splicing proteins, had members of the RENT complex among their PPI interactors. This association had not been previously reported.

While there were complexes with a wide range of functionalities in the splicing subnetwork, most identified subunits were part of the spliceosome (Figure 6E). We built a network for each annotated spliceosome complex shared by more than one bait (Figure 7; Supplementary Table S19). Within these baits, 16 of the 21 members of the commitment complex (or E complex), were detected (61). The three baits with the largest number of E complex subunits were Snp1, Sto1 and Cbc2. Snp1 is a portion of U1 small nuclear RNP (snRNP), and Sto1 and Cbc2 are capping proteins, and are all part of the *S. cerevisiae* commitment complex (59,62). This demonstrates that our approach is able to identify known complexes. The U2-associated complex SF3a/b had a high average coverage ratio across many baits (ratio = 0.67 and 0.83, respectively). However, the other U2 snRNP-associated complex, called RES, was exclusively found in the interactome of its complex member and bait RBP, Ist3.

The U4/U6 × U5 tri-snRNP complex (ratio = 0.82) joins splicing complex A (including the U1 and U2 snRNP complex) to form the preB splicing complex. However, despite the thorough U6 coverage (ratio = 1.0), its complex members were found only with Lsm2 and Dhh1. Lsm2 is a well known complex member of U6 snRNP, whereas Dhh1 has not been characterized in detail as a U6 snRNP-associated protein. Dhh1 facilitates decapping and inhibits translation (63,64). Nevertheless, there has been a yeast two-hybrid study confirming the Lsm2 and Dhh1 interaction (65), and there have been studies associating Lsm4 and Dhh1 with P-granules, which in turn are associated with inhibition of translation (66,67). We noticed that Dhh1 association was limited to snRNPs that join the spliceosome later during the splicing process.

Another unexpected interaction occurred between the RES complex subunit Ist3 and the U5 snRNP complex. The RES complex is critical for the successful formation

of the pre-spliceosome complex via its interaction with U2 snRNP (3). Within the diverse roles of the RES complex, no association with the U5 complex apart from Prp8 was described (63,64). However, within our network analysis we found several PPIs between Ist3 and the U5 snRNP complex members Prp8, Brr2 and Snu114. These data suggest that there might be more U5 members serving as a bridge to the RES complex.

In summary, we found complexes that were critical to the overall mRNA life cycle, with particular enrichment of complexes needed for the assembly and catalytic activity of the spliceosome with putative new roles for RBPs based on concurrent binding patterns.

CONCLUSIONS

Here, we provide an extensive *S. cerevisiae* RBP interactome network to systematically map both PPIs and RDIs. The approach to study RDIs at a larger scale gives the unique opportunity to group RBPs by concurrent binding patterns and thus provides suggestions for functions for the RBPs themselves as well as for their interaction partners. An additional integration of the RINE resource with next-generation sequencing data containing information about the RNAs bound to RBPs would provide further insights into the functionalities of both the RBPs and their associated RNAs. By providing interactive and visual access to the data of this study, the RINE resource (<https://www.butterlab.org/RINE>) can serve as a starting point for further data analysis and exploration of individual candidates.

DATA AVAILABILITY

The R scripts required for the data statistical analysis and its visualization are available at Zenodo (<https://doi.org/10.5281/zenodo.7753608>). The MS proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (68) partner repository with the dataset identifiers PXD035979 (RBP interactome screen) and PXD035971 (KO screen). The protein interactions from this publication have been submitted to the IMEx (<http://www.imexconsortium.org>) consortium through IntAct (69) and assigned the identifier IM-29638.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Assistance by the Proteomics Core Facility and the Media lab at IMB, and critical reading of the manuscript by Julian König is gratefully acknowledged.

FUNDING

The work was supported by the Deutsche Forschungsgemeinschaft [BU 2996/7-1 (SPP1935: ‘Deciphering the mRNP code: RNA-bound Determinants of Post-Transcriptional Gene Regulation’) and project number

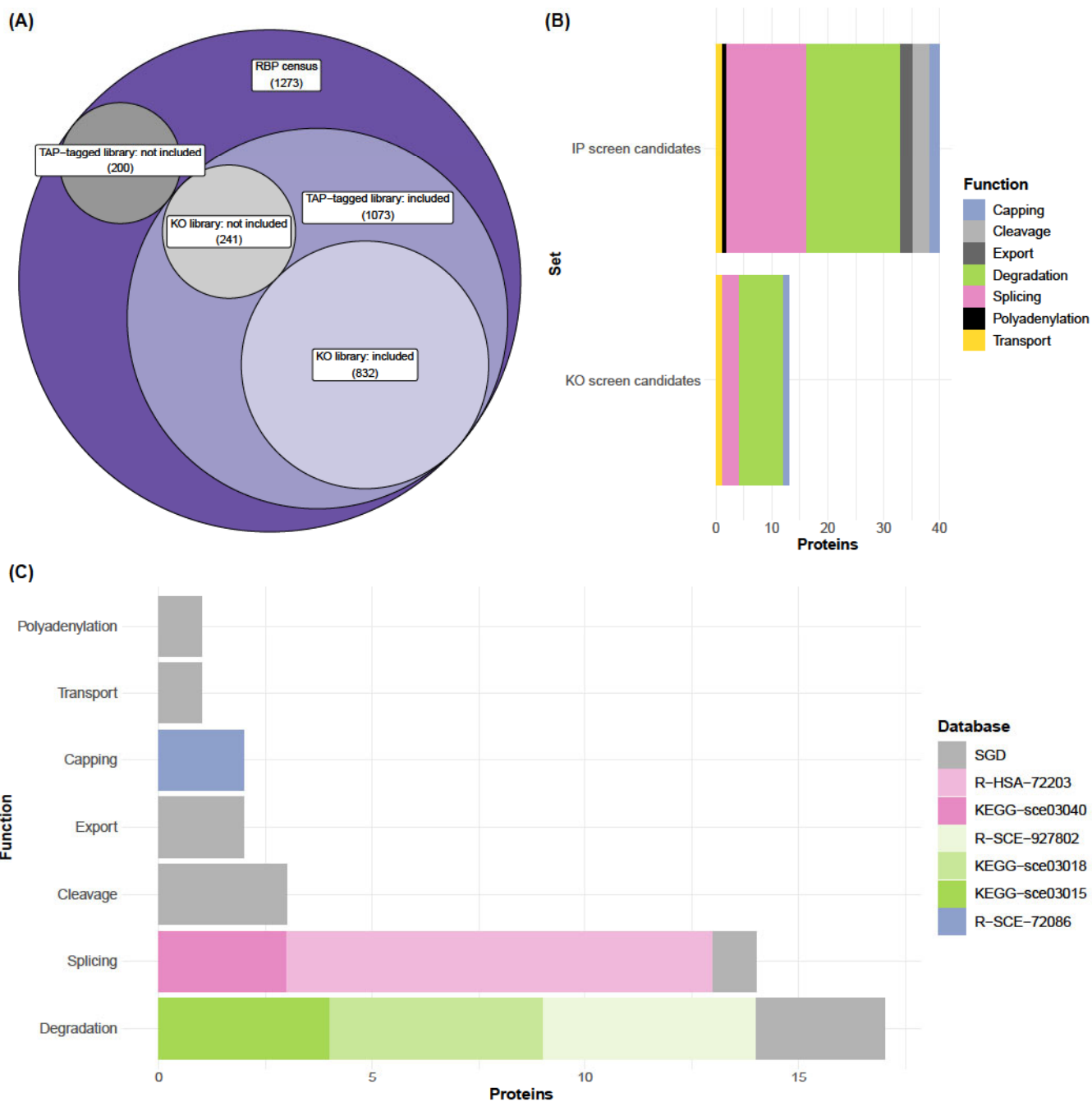
439669440 (TRR319 RMaP TP C03) to F.B. and LU 2568/1-1 to K.L.]. Funding for open access charge: DFG 439669440 [TRR319 RMaP].

Conflict of interest statement. None declared.

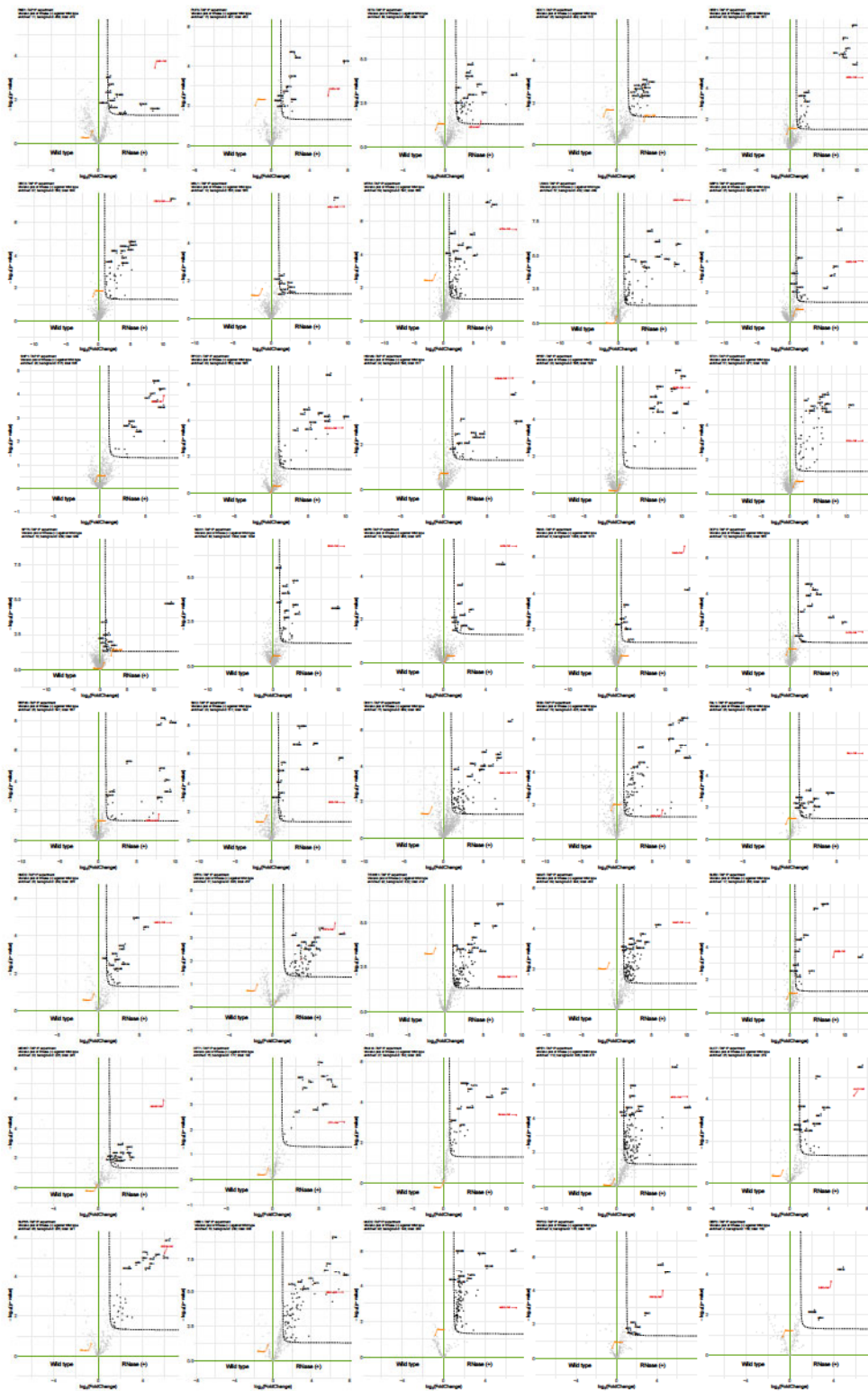
REFERENCES

- Dreyfuss, G., Kim, V.N. and Kataoka, N. (2002) Messenger-RNA-binding proteins and the messages they carry. *Nat. Rev. Mol. Cell Biol.*, **3**, 195–205.
- Ramanathan, A., Robb, G.B. and Chan, S.-H. (2016) mRNA capping: biological functions and applications. *Nucleic Acids Res.*, **44**, 7511–7526.
- Wilkinson, M.E., Charenton, C. and Nagai, K. (2020) RNA splicing by the spliceosome. *Annu. Rev. Biochem.*, **89**, 359–388.
- Neve, J., Patel, R., Wang, Z., Louey, A. and Furger, A.M. (2017) Cleavage and polyadenylation: ending the message expands gene regulation. *RNA Biol.*, **14**, 865–890.
- Tutucci, E. and Stutz, F. (2011) Keeping mRNPs in check during assembly and nuclear export. *Nat. Rev. Mol. Cell Biol.*, **12**, 377–384.
- Xie, Y. and Ren, Y. (2019) Mechanisms of nuclear mRNA export: a structural perspective. *Traffic*, **20**, 829–840.
- Licatalosi, D.D. and Darnell, R.B. (2010) RNA processing and its regulation: global insights into biological networks. *Nat. Rev. Genet.*, **11**, 75–87.
- Mangus, D.A., Evans, M.C. and Jacobson, A. (2003) Poly(A)-binding proteins: multifunctional scaffolds for the post-transcriptional control of gene expression. *Genome Biol.*, **4**, 223.
- Kühn, U. and Wahle, E. (2004) Structure and function of poly(A) binding proteins. *Biochim. Biophys. Acta*, **1678**, 67–84.
- Glisovic, T., Bachorik, J.L., Yong, J. and Dreyfuss, G. (2008) RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.*, **582**, 1977–1986.
- Rissland, O.S. (2017) The organization and regulation of mRNA–protein complexes. *Wiley Interdiscip. Rev. RNA*, **8**, e1369.
- 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.
- 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.
- Mitchell, S.F., Jain, S., She, M. and Parker, R. (2013) Global analysis of yeast mRNPs. *Nat. Struct. Mol. Biol.*, **20**, 127–133.
- Matia-González, A.M., Laing, E.E. and Gerber, A.P. (2015) Conserved mRNA-binding proteomes in eukaryotic organisms. *Nat. Struct. Mol. Biol.*, **22**, 1027–1033.
- Beckmann, B.M., Horos, R., Fischer, B., Castello, A., Eichelbaum, K., Alleaume, A.-M., Schwarzl, T., Curk, T., Foehr, S., Huber, W. et al. (2015) The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. *Nat. Commun.*, **6**, 10127.
- Kramer, K., Sachsenberg, T., Beckmann, B.M., Qamar, S., Boon, K.-L., Hentze, M.W., Kohlbacher, O. and Urlaub, H. (2014) Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. *Nat. Methods*, **11**, 1064–1070.
- Hentze, M.W., Castello, A., Schwarzl, T. and Preiss, T. (2018) A brave new world of RNA-binding proteins. *Nat. Rev. Mol. Cell Biol.*, **19**, 327–341.
- Hentze, M.W. and Preiss, T. (2010) The REM phase of gene regulation. *Trend. Biochem. Sci.*, **35**, 423–426.
- Curtis, N.J. and Jeffery, C.J. (2021) The expanding world of metabolic enzymes moonlighting as RNA binding proteins. *Biochem. Soc. Trans.*, **49**, 1099–1108.
- Singh, R. and Green, M.R. (1993) Sequence-specific binding of transfer RNA by glyceraldehyde-3-phosphate dehydrogenase. *Science*, **259**, 365–368.
- Hentze, M.W. and Argos, P. (1991) Homology between IRE-BP, a regulatory RNA-binding protein, aconitase, and isopropylmalate isomerase. *Nucleic Acids Res.*, **19**, 1739–1740.
- Rouault, T.A., Stout, C., Kaptain, S., Harford, J.B. and Klausner, R.D. (1991) Structural relationship between an iron-regulated RNA-binding protein (IRE-BP) and aconitase: functional implications. *Cell*, **64**, 881–883.
- Kilchert, C., Sträßer, K., Kunetsky, V. and Änkö, M.-L. (2020) From parts lists to functional significance—RNA–protein interactions in gene regulation. *Wiley Interdiscip. Rev. RNA*, **11**, e1582.
- Butter, F., Scheibe, M., Mörl, M. and Mann, M. (2009) Unbiased RNA–protein interaction screen by quantitative proteomics. *Proc. Natl Acad. Sci. USA*, **106**, 10626–10631.
- Scheibe, M., Butter, F., Hafner, M., Tuschl, T. and Mann, M. (2012) Quantitative mass spectrometry and PAR-CLIP to identify RNA–protein interactions. *Nucleic Acids Res.*, **40**, 9897–9902.
- Casas-Vila, N., Sayols, S., Pérez-Martínez, L., Scheibe, M. and Butter, F. (2020) The RNA fold interactome of evolutionary conserved RNA structures in *S. cerevisiae*. *Nat. Commun.*, **11**, 2789.
- Hogan, D.J., Riordan, D.P., Gerber, A.P., Herschlag, D. and Brown, P.O. (2008) Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biol.*, **6**, e255.
- Klass, D.M., Scheibe, M., Butter, F., Hogan, G.J., Mann, M. and Brown, P.O. (2013) Quantitative proteomic analysis reveals concurrent RNA–protein interactions and identifies new RNA-binding proteins in *Saccharomyces cerevisiae*. *Genome Res.*, **23**, 1028–1038.
- Brannan, K.W., Jin, W., Huelga, S.C., Banks, C.A., Gilmore, J.M., Florens, L., Washburn, M.P., Van Nostrand, E.L., Pratt, G.A., Schwinn, M.K. et al. (2016) SONAR discovers RNA-binding proteins from analysis of large-scale protein–protein interactomes. *Mol. Cell*, **64**, 282–293.
- Ghaemmaghami, S., Huh, W.-K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O’Shea, E.K. and Weissman, J.S. (2003) Global analysis of protein expression in yeast. *Nature*, **425**, 737–741.
- Winzler, E.A., Shoemaker, D.D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J.D., Bussey, H. et al. (1999) Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science*, **285**, 901–906.
- Rappsilber, J., Mann, M. and Ishihama, Y. (2007) Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using stagetips. *Nat. Protoc.*, **2**, 1896–1906.
- Cox, J., Neuhauser, N., Michalski, A., Scheltema, R.A., Olsen, J.V. and Mann, M. (2011) Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.*, **10**, 1794–1805.
- Cox, J. and Mann, M. (2008) MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.*, **26**, 1367–1372.
- Cherry, J.M., Hong, E.L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E.T., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S.R. et al. (2012) Saccharomyces genome database: the genomics resource of budding yeast. *Nucleic Acids Res.*, **40**, D700–D705.
- R Core Team (2021) In: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria.
- Wickham, H. (2016) In: *Ggplot2—Elegant Graphics for Data Analysis*. Springer International Publishing, Cham.
- Oughtred, R., Rust, J., Chang, C., Breitkreutz, B.-J., Stark, C., Willems, A., Boucher, L., Leung, G., Kolas, N., Zhang, F. et al. (2021) The BioGRID Database: a comprehensive biomedical resource of curated protein, genetic, and chemical interactions. *Prot. Sci.*, **30**, 187–200.
- 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 Res.*, **49**, D344–D354.
- Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G.A., Sonnhammer, E.L.L., Tosatto, S.C.E., Paladín, L., Raj, S., Richardson, L.J. et al. (2021) Pfam: the protein families database in 2021. *Nucleic Acids Res.*, **49**, D412–D419.
- Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., Chothia, C. and Gough, J. (2009) SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res.*, **37**, D380–D386.
- The Gene Ontology Consortium. (2021) The Gene Ontology resource: enriching a GOLD mine. *Nucleic Acids Res.*, **49**, D325–D334.
- Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L. et al. (2021) clusterProfiler 4.0: a universal enrichment tool for interpreting omics data. *Innovation*, **2**, 100141.
- Kanehisa, M. (2019) Toward understanding the origin and evolution of cellular organisms. *Prot. Sci.*, **28**, 1947–1951.

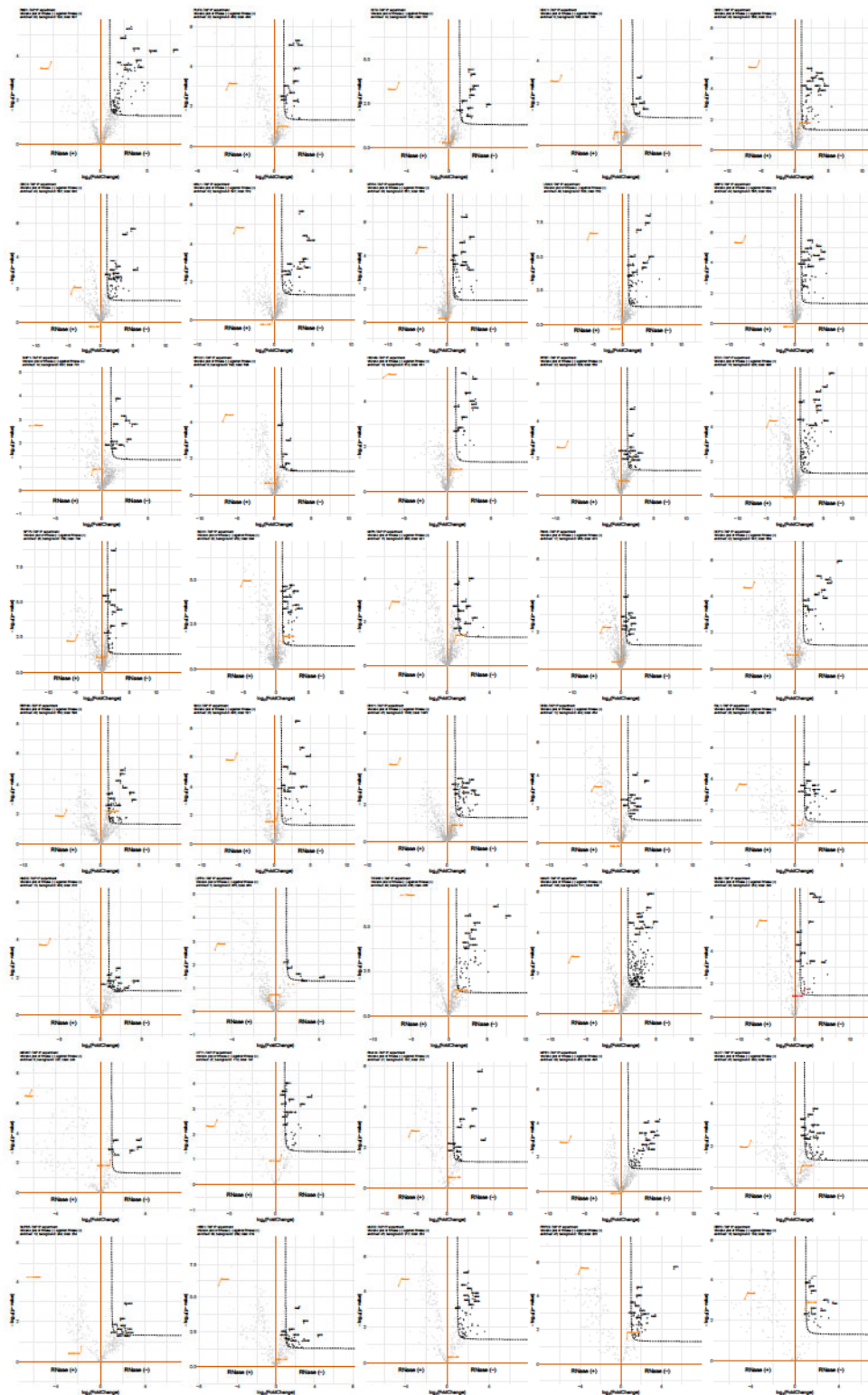
46. Jassal,B., Matthews,L., Viteri,G., Gong,C., Lorente,P., Fabregat,A., Sidiropoulos,K., Cook,J., Gillespie,M., Haw,R. *et al.* (2020) The reactome pathway knowledgebase. *Nucleic Acids Res.*, **48**, D498–D503.
47. Pu,S., Wong,J., Turner,B., Cho,E. and Wodak,S.J. (2009) Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Res.*, **37**, 825–831.
48. Csardi,G. and Nepusz,T. (2006) The igraph software package for complex network research. *InterJournal, Complex Systems*, **1695**, 1–9.
49. Chang,W., Cheng,J., Allaire,J., Sievert,C., Schloerke,B., Xie,Y., Allen,J., McPherson,J., Dipert,A. and Borges,B. (2023) shiny: Web Application Framework for R. R package version 1.7.4.9002, <https://shiny.rstudio.com/>.
50. Ho,Y., Gruhler,A., Heilbut,A., Bader,G.D., Moore,L., Adams,S.-L., Millar,A., Taylor,P., Bennett,K., Boutilier,K. *et al.* (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, **415**, 4.
51. Krogan,N.J., Cagney,G., Yu,H., Zhong,G., Guo,X., Ignatchenko,A., Li,J., Pu,S., Datta,N., Tikuisis,A.P. *et al.* (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*, **440**, 637–643.
52. Corley,M., Burns,M.C. and Yeo,G.W. (2020) How RNA-binding proteins interact with RNA: molecules and mechanisms. *Mol. Cell*, **78**, 9–29.
53. de Godoy,L.M.F., Olsen,J.V., Cox,J., Nielsen,M.L., Hubner,N.C., Fröhlich,F., Walther,T.C. and Mann,M. (2008) Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature*, **455**, 1251–1254.
54. Öztürk,M., Freiwald,A., Cartano,J., Schmitt,R., Dejung,M., Luck,K., Al-Sady,B., Braun,S., Levin,M. and Butter,F. (2022) Proteome effects of genome-wide single gene perturbations. *Nat. Commun.*, **13**, 6153.
55. Bergkessel,M. and Reese,J.C. (2004) An essential role for the *Saccharomyces cerevisiae* DEAD-Box Helicase DHH1 in G1/S DNA-damage checkpoint recovery. *Genetics*, **167**, 21–33.
56. Westmoreland,T., Olson,J., Saito,W., Huper,G., Marks,J. and Bennett,C. (2003) Dhh1 regulates the G1/S-checkpoint following DNA damage or BRCA1 expression in yeast1. *J. Surg. Res.*, **113**, 62–73.
57. Dutertre,M., Lambert,S., Carreira,A., Amor-Guéret,M. and Vagner,S. (2014) DNA damage: RNA-binding proteins protect from near and far. *Trend. Biochem. Sci.*, **39**, 141–149.
58. Klaric,J.A., Wüst,S. and Panier,S. (2021) New faces of old friends: emerging new roles of RNA-binding proteins in the DNA double-strand break response. *Front. Mol. Biosci.*, **8**, 668821.
59. Plaschka,C., Newman,A.J. and Nagai,K. (2019) Structural basis of nuclear pre-mRNA splicing: lessons from yeast. *Cold Spring Harb. Perspect. Biol.*, **11**, a032391.
60. Katahira,J. (2012) mRNA export and the TREX complex. *Biochim. Biophys. Acta*, **1819**, 507–513.
61. Larson,J.D. and Hoskins,A.A. (2017) Dynamics and consequences of spliceosome E complex formation. *eLife*, **6**, e27592.
62. Gonatopoulos-Pournatzis,T. and Cowling,V.H. (2014) Cap-binding complex (CBC). *Biochem. J.*, **457**, 231–242.
63. Sweet,T., Kovalak,C. and Collier,J. (2012) The DEAD-box protein dhh1 promotes decapping by slowing ribosome movement. *PLoS Biol.*, **10**, e1001342.
64. Collier,J.M., Tucker,M., Sheth,U., Valencia-Sanchez,M.A. and Parker,R. (2001) The DEAD box helicase, Dhh1p, functions in mRNA decapping and interacts with both the decapping and deadenylase complexes. *RNA*, **7**, 1717–1727.
65. Uetz,P., Giot,L., Cagney,G., Mansfield,T.A., Judson,R.S., Knight,J.R., Lockshon,D., Narayan,V., Srinivasan,M., Pochart,P. *et al.* (2000) A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, **403**, 623–627.
66. Cary,G.A., Vinh,D. B.N., May,P., Kuestner,R. and Dudley,A.M. (2015) Proteomic analysis of Dhh1 complexes reveals a role for Hsp40 chaperone Ydj1 in yeast P-body assembly. *G3 Genes|Genomes|Genetics*, **5**, 2497–2511.
67. Rao,B.S. and Parker,R. (2017) Numerous interactions act redundantly to assemble a tunable size of P bodies in *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA*, **114**, E9569–E9578.
68. Perez-Riverol,Y., Bai,J., Ban dla,C., García-Seisdedos,D., Hewapathirana,S., Kamatchinathan,S., Kundu,D.J., Prakash,A., Frericks-Zipper,A., Eisenacher,M. *et al.* (2022) The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. *Nucleic Acids Res.*, **50**, D543–D552.
69. Orchard,S., Ammari,M., Aranda.B., Breuza,L., Briganti,L., Broackes-Carter,F., Campbell,N.H., Chavali,G., Chen,C., del-Toro,N. *et al.* (2014) The MIntAct Project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.*, **42**, D358–D363.



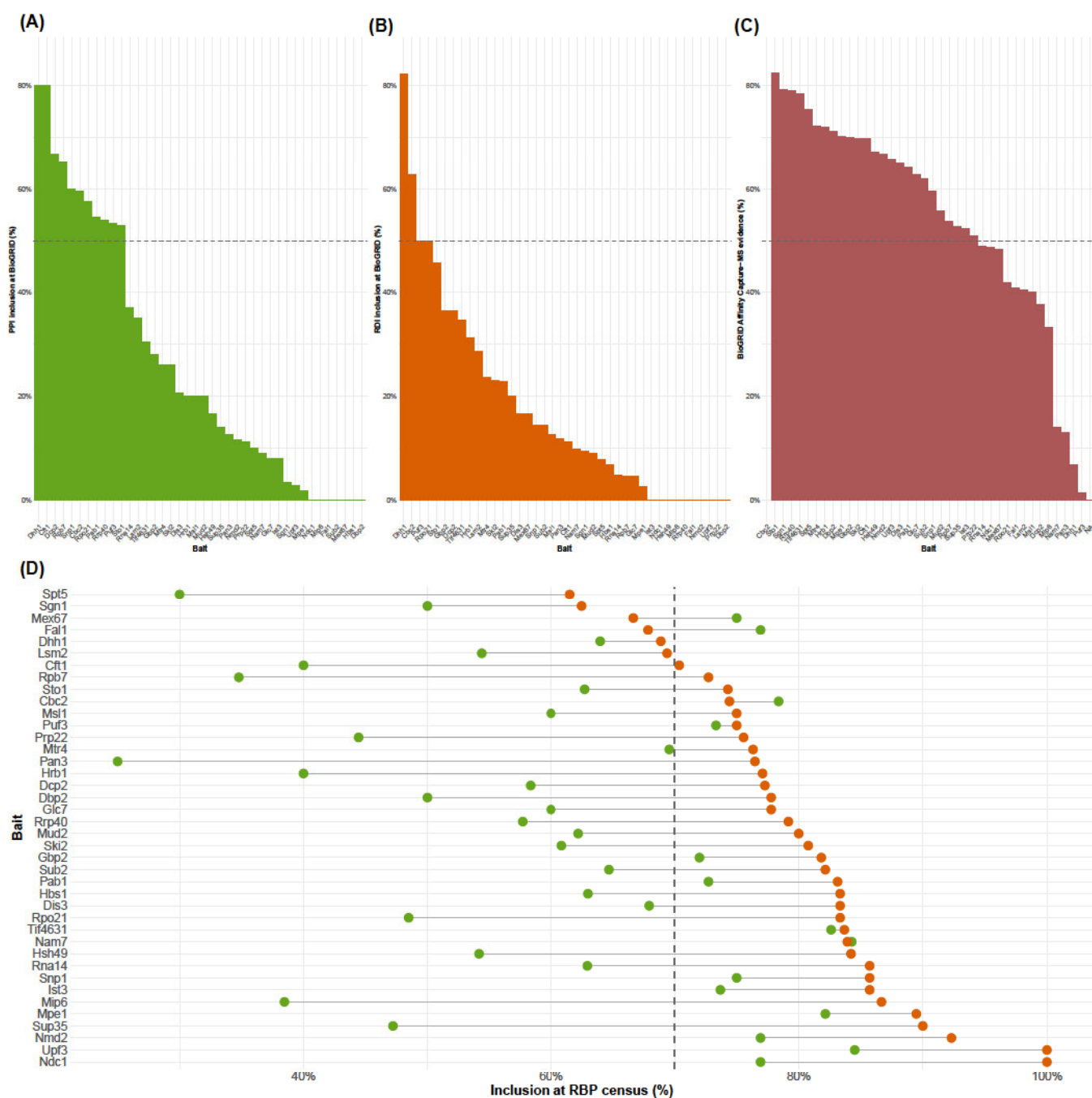
Suppl. Figure 1. Candidate selection for RBP interactome screen. **(A)** Comparison of RBP census with TAP-tagged and KO libraries. Each group is represented by a circle. **(B)** Bar plot of the selected candidates for each screen with their main functional mRNA pathway. **(C)** Barplot counting the interactome screen bait RBPs per major RNA process, differentiating by particular KEGG pathway (KEGG), reactome term (R) and Saccharomyces Genome Database (SGD) entry.



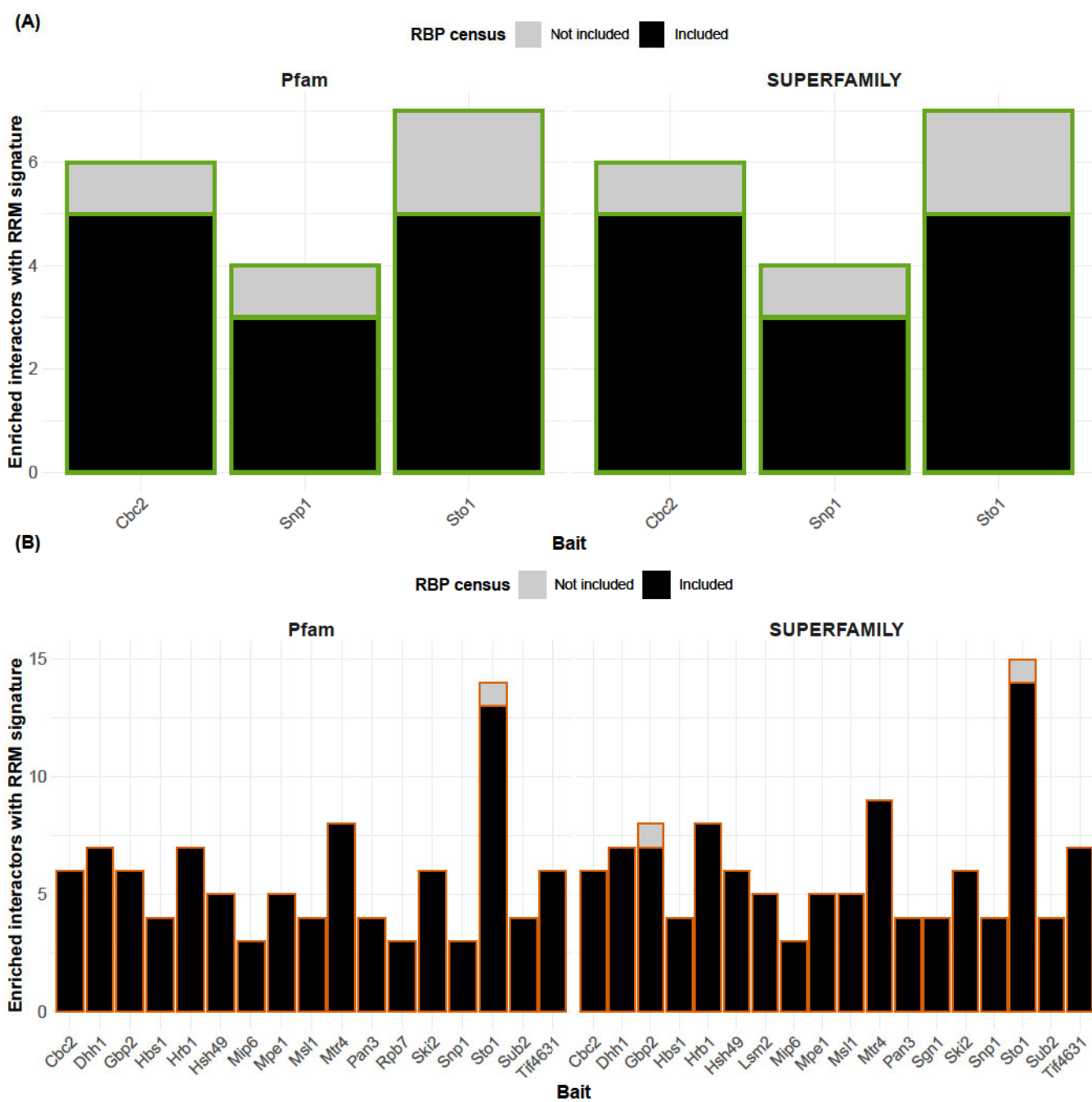
Suppl. Figure 2. PPI volcano plots. Each plot visualises quantified proteins per bait-TAP IP, digested with RNase A, compared to WT without tagged bait determined by quantitative proteomics ($n=4$, for RNA14 $n=3$). Enrichment threshold (dotted line) set to p -value < 0.05 (Welch t -test) and fold change > 2 , $c = 0.05$. Each dot represents a protein with enriched interactors (black), tagged bait-RBP (enriched in red, otherwise orange) and RNase A (orange) highlighted.



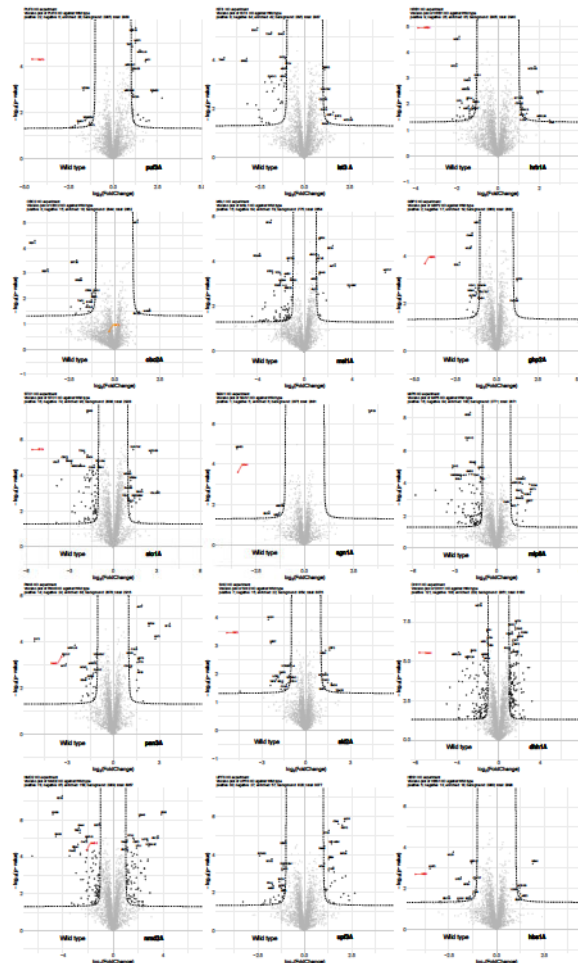
Suppl. Figure 3. RDI volcano plots. Each plot visualises quantified proteins per bait-TAP IP, without RNase A, to bait-TAP IP treated with RNase A determined by quantitative proteomics ($n=4$, for RNA14 $n=3$). Enrichment threshold (dotted line) set to p -value < 0.05 (Welch t-test) and fold change > 2 , $c = 0.05$. Each dot represents a protein with enriched interactors (black), tagged bait-RBP (enriched in red, otherwise orange) and RNase A (orange) highlighted.



Suppl. Figure 4. BioGRID and RBP census analysis. **(A and B)** Barplot indicating the percentage of enriched interactors identified as PPI (A) and RDI (B). **(C)** Ratio of interactors with 'affinity capture-MS' as reported experimental evidence in BioGRID. **(D)** Cleveland dot-plot of enriched interactors identified as PPI (green) and RDI (orange) included in the RBP census.

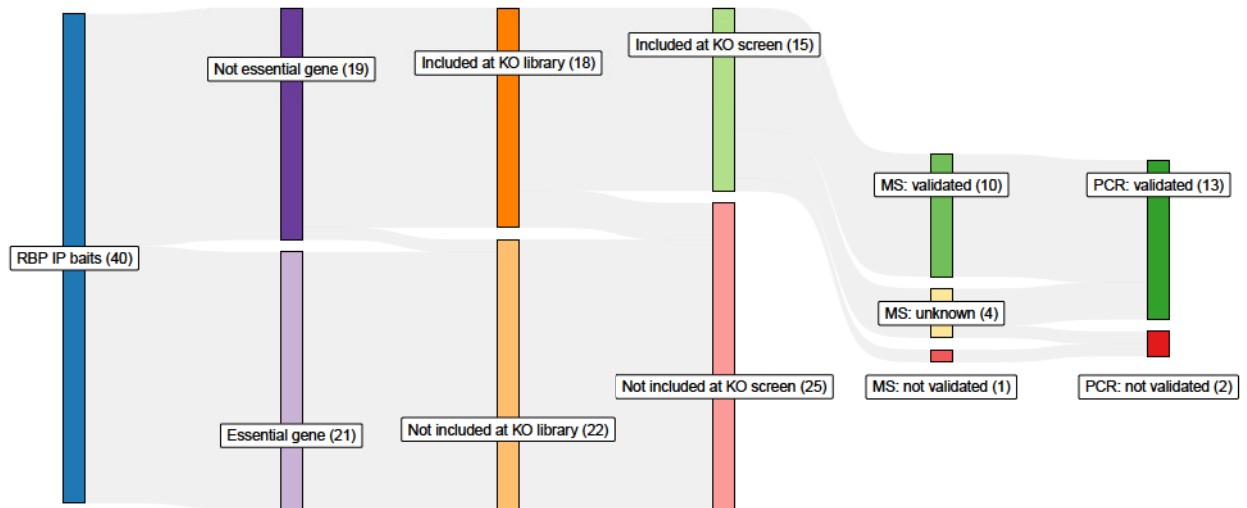


Suppl. Figure 5. RNA recognition motif (RRM) domain representation among PPI and RDI. **(A and B)** Barplot of the number of PPIs (A) and RDIs (B) with the RRM signature overrepresented (p -value < 0.01 ; Fisher's exact test). Interactors included in the RBP census are highlighted (black).

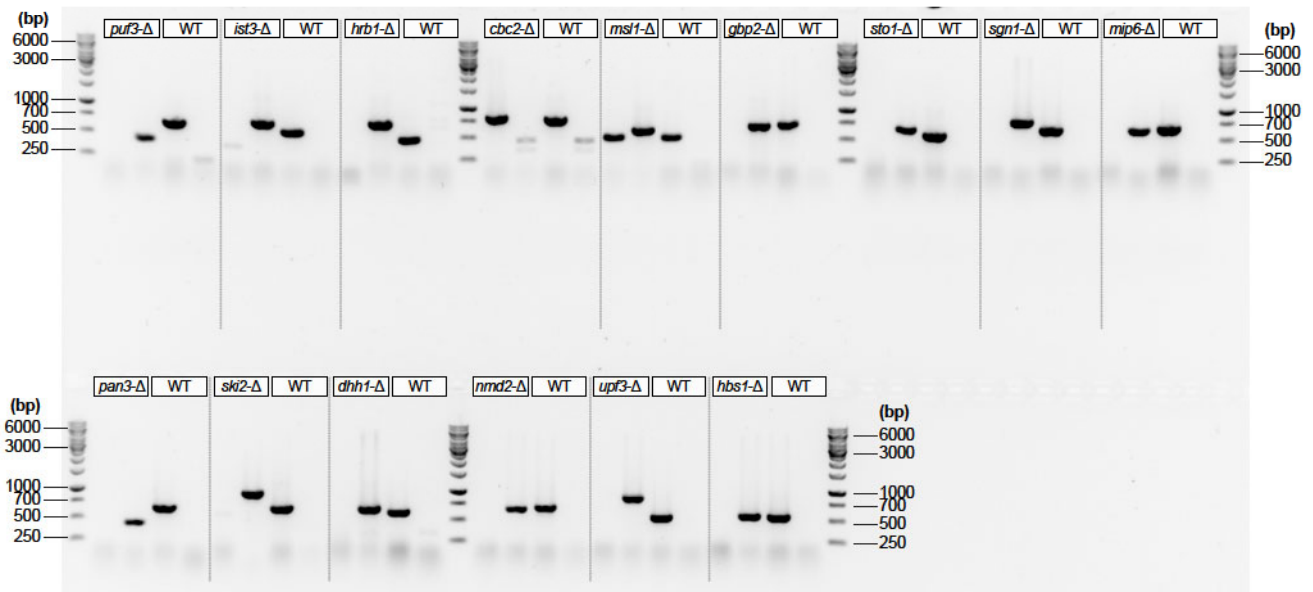


Suppl. Figure 6. KO screen volcano plots. Each plot visualises quantified proteins per RBP KO to WT proteome by quantitative proteomics (n= 4, for *mip6* Δ n =3). Enrichment threshold (dotted line) set to p-value < 0.05 (Welch t-test) and abs(fold change) > 2, c = 0.05. Each dot represents a protein with enriched interactors (black) and marks knocked out RBP (down-regulated red, otherwise orange).

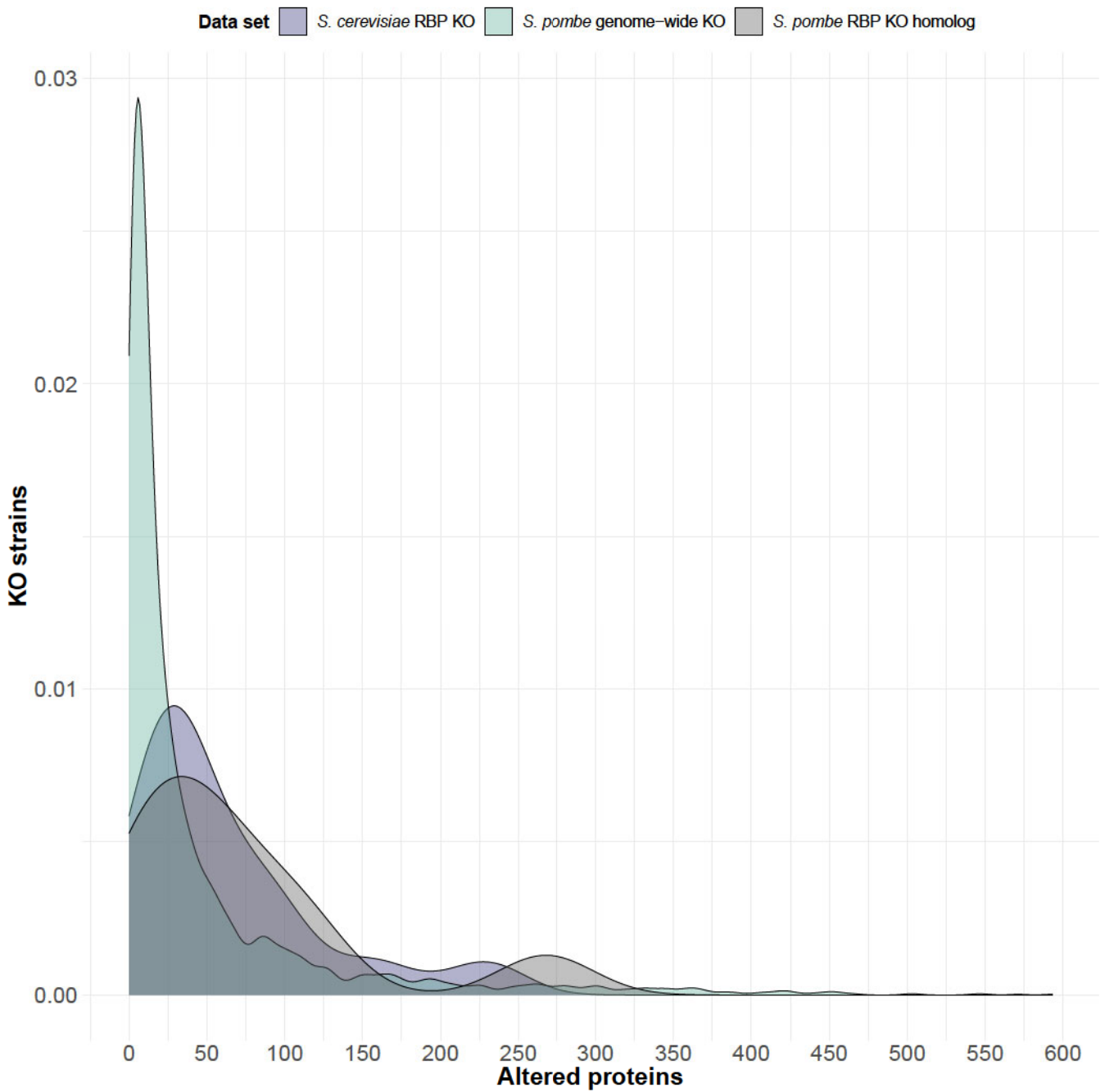
(A)



(B)



Suppl. Figure 7. KO strain selection and PCR validation. (A) Sankey diagram depicting selection of the included KO strains. (B) Agarose gel of PCR validation for KO and WT strains. First lane detected the wild-type allele, second lane the deletion allele.



Suppl. Figure 8. Altered protein distribution in *S. cerevisiae* and *S. pombe*. Density plot visualising the number of altered proteins upon KO for the *S. cerevisiae* RBP KO screen (blue), the *S. pombe* genome-wide KO library (green) and the *S. pombe* RBP KO screen homologues (grey).

4. Conclusions and future perspectives

4.1 Telomeric repeat-containing RNA origin and interacting partners

Differences in the behaviour of TERRA molecules in *H. sapiens* and *M. musculus* were initially observed at co-localization level. While TERRA-FISH patterns were prominently detected at telomeres in *H. sapiens*, such co-localization patterns with telomeres were rarely observed in *M. musculus* (De Silanes et al., 2014). Article I delved further into the investigation of TERRA molecules' genesis and confirmed a distinct genomic origin for *M. musculus* compared to *H. sapiens* (Viceconte et al., 2021).

The transcription origin of TERRA molecules was investigated using a FISH approach. We demonstrated that while TERRA-FISH intensity levels and telomere length were correlated in *H. sapiens*, no such correlation existed in *M. musculus*. The results in *H. sapiens* were consistent with previous northern blot experiments (Arnoult et al., 2012; Van Beneden et al., 2013; Yehezkel et al., 2008) and indicates a telomeric TERRA origin. Conversely, the lack of correlation between intensity levels and telomere length in *M. musculus* suggested an additional genomic origin apart from telomeres. In this context, a previous CHIRT experiment had revealed that the pseudoautosomal PAR locus of *M. musculus* embryonic stem cells (ESC) produced a TERRA-like transcript capable of binding, in trans, to most telomeres (Chu, Froberg, et al., 2017).

Additionally, we observed that overall TERRA transcript levels did not correlate with telomere length in either *H. sapiens* or *M. musculus*. We suggest that in *H. sapiens*, only telomere-bound TERRA molecules are detected with RNA-FISH, while other TERRA molecules remain undetected, explaining the lack of correlation between telomere length and overall TERRA transcript levels. This might be happening due to technical reasons, such as the pre-extraction step of the RNA-FISH protocol, which could wash away unbound RNAs, or the presence of G4 secondary structures forming along the TERRA (UUAGGG)_n repeats (Biffi et al., 2012; Hirashima & Seimiya, 2015) hindering their recognition by the FISH probes. Meanwhile, in *M. musculus*, we demonstrated that the vast majority of UAAGGG repeats originated from the PAR locus, explaining the lack of correlation between overall TERRA transcript levels and

telomere length. Altogether, the RNA-FISH experiments pointed to distinct origins for TERRA molecules in *H. sapiens* and *M. musculus*.

We further investigated the genomic origins of TERRA with a series of *in silico* experiments. Hence, we searched for TERRA loci along the genome using publicly available RNA-Seq datasets. We obtained reads mapping to the already-known Telo 18q and PAR locus locations. The latter is located in an intronic region of the *Erdr1* gene on the X/Y chromosome, reinforcing the idea that spliced introns could be non-coding RNA precursors (Hesselberth, 2013). We also unveiled a novel TERRA locus on chromosome 2, likely located within the 3'UTR region of the *Polr3k* gene. Interestingly, 3'UTR regions have been proposed as lncRNAs templates, either as a whole or as cleaved fragments (Mayr, 2017). This locus contributes significantly to the overall pool of TERRA molecules in some *M. musculus* brain tissues, including the forebrain and the frontal lobe. A deeper analysis of the CHIRT data (Chu, Froberg, et al., 2017) also reveals a TERRA-binding peak at the *Pol3k* locus. This peak likely corresponds to Chr 2 TERRA molecules, suggesting that they bind where they are produced. Finally, we revealed one last intrachromosomal source of (UUAGGG)_n-containing transcripts produced at the *Tsix* locus, which is located within the X-inactivation centre (Xic). This was found exclusively in *M. musculus* ESC. It has been shown that, in ESC, an interaction between the PAR and Xic locus initiates the inactivation of chromosome X (Chu, Froberg, et al., 2017). Within the same investigation, CHIRT data suggested that the *Tsix* locus was bound not only by PAR-TERRA, but also by other (UUAGGG)_n-containing transcripts (Chu, Froberg, et al., 2017). We thus suggest that, in *M. musculus* ESC, the TERRA-like (UUAGGG)_n-containing transcripts produced at the *Tsix* locus may interact with PAR-TERRA and promote Xic::PAR pairing, which, in turn, promotes the inactivation of the X chromosome.

Despite their distinct genomic origins, TERRA molecules seem to play similar roles in *H. sapiens* and *M. musculus* (Bettin et al., 2019). This means that TERRA-like RNAs, transcribed from different loci, bind to various genomic loci, including telomeres, to exert similar functionalities. We hypothesised this might be possible because they share a pool of common interactors. Thus, we designed a comparative MS-SILAC pull-down screen to identify TERRA-interacting proteins in *H. sapiens* and *M. musculus*.

Overall, TERRA interactors found in both species included proteins involved in DNA replication and repair, pre-mRNA processing, rRNA metabolism and Pol II-dependent transcription. These included subunits of the PRC2 complex, HNRNP proteins and the BLM helicase. We also found several Aurora kinases, including the Aurora kinase B (*Aurkb*). Interestingly, in S phase *M. musculus* ESC, *Aurkb* has been reported to recruit telomerase to

chromosome ends through its interaction with centromeric RNAs (Mallm & Rippe, 2015). Thus, Aurkb's proposed role is to enhance the telomerase complex activity through its interaction with centromeric RNAs. Interestingly, a similar function has been observed in *S. cerevisiae* and *S. pombe*, where Aurkb interacts with TERRA to activate telomerase (Cusanelli et al., 2013; Moravec et al., 2016). Additionally, Aurkb has been reported to localise at *M. musculus* ESC telomeres. There, it modulates Terf1's affinity for telomeres, which, in turn, plays a role in telomere integrity (Chan et al., 2017). We thus suggest that, in *M. musculus*, the TERRA-Aurkb interaction might also be implicated in telomere protection by modulating shelterin proteins' affinity for telomeres. Another group of proteins found among the TERRA interactors of both *H. sapiens* and *M. musculus* screens includes components of the paraspeckle subnuclear bodies and the Dazap1 RBP. In *H. sapiens*, the interaction between TERRA and the NONO paraspeckle component has been reported, and it has been shown to suppress RNA:DNA hybrid-induced telomere instability. Within the same study, the interaction between TERRA and DAZAP1 was reported (Petti et al., 2019).

There are two groups of proteins that are unique to each screen. On one hand, several centromeric proteins were found to be interacting with TERRA in *M. musculus*, which suggests a role for TERRA in centromere assembly or stability. On the other hand, several proteins from the exosome complex were pulled-down in the *H. sapiens* screen, but not in *M. musculus*. This might be linked to distinct degradation machineries between species.

We also investigated differences between TERRA and PAR-TERRA interactors in *M. musculus*. To do so, the results from the *M. musculus* TERRA pull-down screen were overlapped with those obtained with an *in vivo* iDRiP (Minajigi et al., 2015) screen in ESC, where PAR-TERRA was likely the most abundant TERRA species (Chu, Cifuentes-Rojas, et al., 2017). The Atrx protein is missing in the TERRA pull-down screen, with respect to the iDRiP screen. This might be related to the length of the TERRA pull-down probe. It has been shown that, while an 83-nt long *in vitro* synthesised TERRA-RNA probe was efficiently shifted by Atrx in an electrophoretic mobility shift assay, a shorter, 30-nt long probe was less efficiently shifted (Chu, Cifuentes-Rojas, et al., 2017). Hence, the (UUAGGG)₈ TERRA-pull-down screen probe might not be long enough to allow an efficient interaction with Atrx. Other missing iDRiP-screen proteins with important roles in telomere biology, such as Rtek1, Rpa1/2, Ctc1, Stn1, or Pml (Chu, Cifuentes-Rojas, et al., 2017), might also be missing because of the probe length. Another plausible explanation would be that an *in vitro* assay such as SILAC pull-down fails to recapitulate the chromatic context. Despite these missing proteins, there is a good overlap between both screens. This might indicate that, whether they originate from the transcription of

telomeres or the intrachromosomal PAR locus, the UUAGGG-rich noncoding RNA molecules interact with the same proteins and are thus likely to exert similar functions.

The investigation presented in Article I (Viceconte et al., 2021) reflects how MS-techniques are currently used for investigating interactions between proteins and non-coding RNA species. In this case, a SILAC label-based MS screen was designed to pull down TERRA interactors. This enabled first insights into TERRA and PAR-TERRA functionalities. Still, future experiments will be necessary to properly address their functionalities, both in *H. sapiens* and *M. musculus*. These might include further MS techniques, such as the assessment of proteome changes after depletion of TERRA and/or PAR-TERRA transcripts.

4.2 RNA binding protein network-based function assignment

Over the last decade, the number of proteins identified as RBPs has greatly increased, and there are currently over 1.000 proteins for the model organisms *H. sapiens*, *M. musculus*, and *S. cerevisiae* (Hentze et al., 2018). Despite this rapid growth, the cellular function for most of these novel RBPs is still to be investigated. Article II focused on providing a framework for novel RBP function assignment, relying on MS-based interactomics (Fradera-Sola et al., 2023).

We present a novel data set for 40 *S. cerevisiae* RBPs, which includes less characterised proteins from recent RIC studies (Beckmann et al., 2015; Matia-González et al., 2015; Mitchell et al., 2013) and that spans over an mRNA life's cycle, from its nuclear processing to its cytoplasmic degradation. The dataset was generated using the RBPs as IP bait and with a label-free MS quantitative interactome screen, as previously shown (Klass et al., 2013). It resulted in two distinct groups of enriched interactors: (i) the PPI group, obtained from RNase-treated IPs, and (ii) the RDI group, obtained from untreated IPs. Within the RDI group, the majority of interactors were included in the RBP census (Hentze et al., 2018), and the proportion of RBPs was higher than in the PPI group. Thus, our approach can unravel hitherto unknown RDIs among a large set of RBPs.

Next, we focused on RNA functionalities and investigated whether they were overrepresented among the enriched interactors from the PPI and RDI groups. We queried them for functional annotations in several databases, from the structural protein domain level, using Pfam and SUPERFAMILY, to the molecular function and pathway level using GO and KEGG, respectively.

At the domain level, we observed a general trend of the RDI group having a larger amount of canonical RBDs, which aligns with the higher proportion of proteins identified as RBPs in the RDI group. Hence, the higher number of RBDs found in the RDI group might be linked to known RBPs. Interestingly, at the molecular function level, both the PPI and RDI group showed an enrichment for terms related to nucleic acid binding, particularly RNA and mRNA binding. Still, the number of baits with overrepresented RNA-related molecular functionalities is higher for the RDI group than for the PPI group. Despite the domain and GO molecular function revealing an enrichment for RNA functionalities, especially for the RDI group, the overall number of canonical RBDs is low. This indicates that RBPs lacking canonical RBDs might be abundant among our enriched interactors and may have less-studied functions in the context of RNA biology.

At the pathway level, we obtained, as expected, several over-represented KEGG terms associated with known bait RBP functionality for both the PPI and RDI groups. Additionally, even though it was not among the selection criteria for bait RBPs, we found a strong overrepresentation of metabolic and synthesis pathways. Metabolic proteins have been proposed to function as RNA binders without canonical RBDs (Hentze et al., 2018), which coincides with the observed lack of RBD enrichment across all bait RBPs. Hence, our dataset provides more evidence for the growing field of metabolic enzymes moonlighting as RBPs (Curtis & Jeffery, 2021; Hentze & Preiss, 2010).

Overall, the RNA-centric functional analysis of the RDI and PPI groups indicates that, despite a high number of interactors without canonical RBDs, the obtained interactors are involved in RNA biology through their association with structural domains, molecular functions, and pathway annotations. Hence, we suggest that functionalities for RBPs lacking annotations could be inferred through their interaction with well-characterised RBPs. A similar concept has been used for predicting hitherto unknown RBPs, as shown for a smaller set of baits in the SONAR dataset (Brannan et al., 2016).

RBPs have been shown to post-transcriptionally regulate genes via their interaction with the transcripts of the regulated genes (Corley et al., 2020; Glisovic et al., 2008). Hence, we wanted to investigate which downstream biological processes are likely to be regulated by the RBP baits screened in our IP-screen. To do so, we aimed to identify differentially expressed proteins upon RBP bait KO using a quantitative label-free MS screen. This KO screen was conducted on a subset of RBPs, composed mainly of non-essential bait RBPs included in the interactome screen. We quantified approximately two-thirds of the expressed *S. cerevisiae* proteome per KO strain (de Godoy et al., 2008) and obtained, on average, a higher number of

differentially expressed proteins than those observed in a genome-wide KO screen performed on *S. pombe* (Öztürk et al., 2022). This indicates that, on average, the RBP-KO led to more profound expression changes. The limited overlap of these differentially expressed proteins with the PPI and RDI sets of each respective bait RBP supports the idea that these RBPs regulate downstream processes.

The KO screen functional analysis mostly revealed amino acid and nucleic acid synthesis, as well as several metabolic pathways that were overrepresented. These pathways did not overlap with the overrepresented pathways found in the RBP interactome screen. This lack of overlap aligns with the previously observed limited overlap between the interactome and KO screens, further indicating the regulation of downstream processes. Additionally, we annotated the differentially expressed proteins with the manually curated protein complexes included in the CYC2008 dataset (Pu et al., 2009). We primarily obtained dimeric complexes covering a wide range of functionalities, some of which mirrored the overrepresented KEGG pathways. For example, we found several metabolism-related complexes up-regulated in *dhh1*- Δ . Similarly, *mip6*- Δ and *gbp2*- Δ had a subunit of the Pmt3p/Pmt5p metabolic complex downregulated. We also identified deregulated complexes not captured within the KEGG analysis. Among the down-regulated proteins in *dhh1*- Δ , there are proteins involved in mismatch repair, aligning with emerging studies linking RBPs with DNA repair processes (Dutertre et al., 2014; Klaric et al., 2021). Furthermore, Dhh1 has been reported to be involved in DNA repair (Bergkessel & Reese, 2004) and to play a critical role in the G₁/S phase cell cycle; its deletion leads to sensitivity to ionising radiation (Westmoreland et al., 2003). Collectively, the KO screen results suggest possible downstream processes regulated by the selected RBPs and could lead to further investigations of RBPs' roles in DNA repair and metabolism. The differentially expressed proteins from the KO screen could guide future experimental investigations to unravel the roles of RBPs in these pathways.

To provide novel functional annotations for the investigated RBPs, we used the PPI and RDI groups from the interactome screen to build an extensive interaction network. In each network, we included only bait RBPs with overrepresented KEGG pathways. This allowed us to visualise broader functionalities and to create specific function-based subnetworks. These subnetworks were further annotated with the protein complexes included in the CYC2008 dataset. Together, the function-based subnetworks and the complex annotation provide a unique framework for examining cross-talk between complex members and inferring novel RBPs' functionalities. This concept is demonstrated in Article II for splicing, an exemplary RNA process widely studied in *S. cerevisiae* (Plaschka et al., 2019). Within the PPI and RDI

subnetworks, we found 19 complexes with more than half of their members in the networks, including the spliceosome components. Other complexes involved in the mRNA life cycle, such as capping, nuclear export, or degradation, were also found. For example, members of the transcription export complex (TREX) and the THO complex, critical for the nuclear export of mRNA (Katahira, 2012), were found to interact with six bait RBPs.

As anticipated, most of the identified complexes were subunits of the spliceosome and previously reported interactions. For instance, 16 of the 21 members of the commitment complex (or E complex) (Larson & Hoskins, 2017) were found within the PPI and RDI networks, primarily in the interactions of the Sto1, Cbc2 and Snp1 bait RBPs. Sto1 and Cbc2 are capping proteins, while Snp1 is a component of the U1 small nuclear RNP (snRNP), all of which are part of the *S. cerevisiae* commitment complex (Gonatopoulos-Pournatzis & Cowling, 2014; Plaschka et al., 2019). Another known spliceosome-associated complex with a high coverage across many bait RBPs is the SF3a/b complex, which interacts with the U2 subunit. However, RES, the other complex known to interact with the U2 snRNP, was exclusively found in the interactome of its complex member and bait RBP, Ist3. Overall, this demonstrates that within the RNA-dependent RBP interactome framework, we can isolate specific functionalities and identify their associated known complexes.

In addition to identifying known interactions and complexes, we were also able to unravel potential new functionalities for the selected bait RBPs. For example, complex members of the U6 snRNP, part of the larger U4/U6 x U5 tri-snRNP complex, were found to be interacting with Lsm2 and Dhh1. While Lsm2 is a well-known complex member of the U6 snRNP, Dhh1 has not been extensively characterised as part of the complex. Dhh1 has been shown to facilitate decapping and inhibit translation (Sweet et al., 2012; Coller et al., 2001). It has also been shown to interact with Lsm2 in a yeast two-hybrid study (Uetz et al., 2000) and, together with Lsm4, has been associated with P-granules, which are, in turn, associated with the inhibition of translation (Cary et al., 2015; Rao & Parker, 2017). Thus, we noticed that Dhh1 interactions are limited to snRNPs that join the spliceosome complex in the later stages of splicing.

Still within the spliceosome, we found an unexpected interaction between the RES complex subunit Ist3 and the U5 snRNP. The RES complex is known to interact with the U2 snRNP and plays a critical role in the formation of the pre-spliceosome (Wilkinson et al., 2020). Among the diverse roles of the RES complex, no association with U5 snRNP, apart from Prp8, has been described (Galej et al., 2013; Schneider et al., 2015). However, we identified PPIs between Ist3 and the U5 snRNP complex members Prp8, Brr2, and Snu114 within our network

analysis. Thus, the data suggests that there might be more interactions serving as a bridge between the U5 snRNP and the RES complex than previously thought.

In summary, within the RBP network analysis, we identified known complexes critical to the overall mRNA life cycle. This knowledge, combined with the identification of concurrent binding patterns, allows us to suggest putative new roles for the studied RBPs. This is shown for splicing with several novel interactions between RBPs and the spliceosome complexes.

The investigation presented in Article II (Fradera-Sola et al., 2023) provides an extensive *S. cerevisiae* RNA-dependent RBP interactome network. Within this network, RBPs are categorised based on their concurrent binding patterns and functionalities. This framework offers a unique avenue for proposing new functionalities not only for the RBPs themselves but also for their interacting partners. The data is available at the RBP interactive network explorer (RINE) at <https://www.butterlab.org/RINE/>. By granting visual and interactive access to this data, we aim for the RINE resource to serve as an initial stepping stone for further data analysis and comprehensive characterization of individual candidates.

4.3 MS-based proteomics to study RNA-protein interactions

This thesis has delved into the study of RNA-protein interactions using MS-based proteomics and interactomics techniques. Thus, we have gained profound insights into the intricate interplay between RNA molecules and proteins. MS-based proteomics and interactomics have proven to be invaluable tools in deciphering the complexity of RNA-protein interactions. These methods allow us to identify and characterise a plethora of RBPs, revealing their context-specific roles in various cellular compartments and conditions. Our research, presented in Article I (Viceconte et al., 2021) and Article II (Fradera-Sola et al., 2023), has highlighted the critical roles that non-coding and coding RNA species, respectively, play in cellular processes through their interaction with proteins.

To obtain a more comprehensive understanding of RBP function, integrating MS-based proteomics with other omics technologies, such as NGS-based transcriptomics and epigenomics, is essential. This multi-omics approach will enable researchers to analyse RNA-protein interactions in the context of gene expression and chromatin modifications, providing a more holistic view of post-transcriptional regulation.

Other avenues worth exploring in the field also include spatial proteomics and single-molecule resolution techniques. With spatial proteomics, one can gain insights into the subcellular localization of RBPs and their interacting RNA partners, offering a unique opportunity

to unravel the spatial dynamics of RNA-protein interactions. With single-cell proteomics, one can gain insights into the heterogeneity of interactions and unravel transient binding events. Single-cell techniques can provide unprecedented insights into the kinetics and stoichiometry of RNA-protein interactions, offering a deeper understanding of their functional significance.

In summary, the exploration of RNA-protein interactions through mass spectrometry-based proteomics and interactomics has provided a deeper understanding of the studied cellular processes. As technology continues to advance, embracing single-molecule, spatial, and multi-omics approaches will undoubtedly unravel novel layers of complexity in RNA-protein networks.

5. References

- Aebersold, R., & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, *422*(6928), 198–207. <https://doi.org/10.1038/nature01511>
- Aebersold, R., & Mann, M. (2016). Mass-spectrometric exploration of proteome structure and function. *Nature*, *537*(7620), 347–355. <https://doi.org/10.1038/nature19949>
- Ahmed, F. E. (2009). Sample preparation and fractionation for proteome analysis and cancer biomarker discovery by mass spectrometry. *Journal of Separation Science*, NA-NA. <https://doi.org/10.1002/jssc.200800622>
- Al Shweiki, M. R., Mönchgesang, S., Majovsky, P., Thieme, D., Trutschel, D., & Hoehenwarter, W. (2017). Assessment of Label-Free Quantification in Discovery Proteomics and Impact of Technological Factors and Natural Variability of Protein Abundance. *Journal of Proteome Research*, *16*(4), 1410–1424. <https://doi.org/10.1021/acs.jproteome.6b00645>
- Allet, N., Barrillat, N., Baussant, T., Boiteau, C., Botti, P., Bougueleret, L., Budin, N., Canet, D., Carraud, S., Chiappe, D., Christmann, N., Colinge, J., Cusin, I., Dafflon, N., Depresle, B., Fasso, I., Frauchiger, P., Gaertner, H., Gleizes, A., ... Zwahlen, C. (2004). In vitro and in silico processes to identify differentially expressed proteins. *PROTEOMICS*, *4*(8), 2333–2351. <https://doi.org/10.1002/pmic.200300840>
- Arnoult, N., Van Beneden, A., & Decottignies, A. (2012). Telomere length regulates TERRA levels through increased trimethylation of telomeric H3K9 and HP1 α . *Nature Structural & Molecular Biology*, *19*(9), 948–956. <https://doi.org/10.1038/nsmb.2364>
- Azzalin, C. M., Reichenbach, P., Khoriantuli, L., Giulotto, E., & Lingner, J. (2007). Telomeric Repeat-Containing RNA and RNA Surveillance Factors at Mammalian Chromosome Ends. *Science*, *318*(5851), 798–801. <https://doi.org/10.1126/science.1147182>
- Baltz, A. G., Munschauer, M., Schwanhäusser, B., Vasile, A., Murakawa, Y., Schueler, M., Youngs, N., Penfold-Brown, D., Drew, K., Milek, M., Wyler, E., Bonneau, R., Selbach, M., Dieterich, C., & Landthaler, M. (2012). The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts. *Molecular Cell*, *46*(5), 674–690. <https://doi.org/10.1016/j.molcel.2012.05.021>
- Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., & Kuster, B. (2007). Quantitative mass spectrometry in proteomics: A critical review. *Analytical and Bioanalytical Chemistry*, *389*(4), 1017–1031. <https://doi.org/10.1007/s00216-007-1486-6>
- Beckmann, B. M., Horos, R., Fischer, B., Castello, A., Eichelbaum, K., Alleaume, A.-M., Schwarzl, T., Curk, T., Foehr, S., Huber, W., Krijgsveld, J., & Hentze, M. W. (2015). The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. *Nature Communications*, *6*(1), 10127. <https://doi.org/10.1038/ncomms10127>
- Bergkessel, M., & Reese, J. C. (2004). An Essential Role for the *Saccharomyces cerevisiae* DEAD-Box Helicase DHH1 in G1/S DNA-Damage Checkpoint Recovery. *Genetics*, *167*(1), 21–33. <https://doi.org/10.1534/genetics.167.1.21>
- Bettin, N., Oss Pegorar, C., & Cusanelli, E. (2019). The Emerging Roles of TERRA in Telomere Maintenance and Genome Stability. *Cells*, *8*(3), 246. <https://doi.org/10.3390/cells8030246>

- Biemann, K. (1992). MASS SPECTROMETRY OF PEPTIDES AND PROTEINS. *Annual Review of Biochemistry*, 61(1), 977–1010. <https://doi.org/10.1146/annurev.bi.61.070192.004553>
- Biffi, G., Tannahill, D., & Balasubramanian, S. (2012). An Intramolecular G-Quadruplex Structure Is Required for Binding of Telomeric Repeat-Containing RNA to the Telomeric Protein TRF2. *Journal of the American Chemical Society*, 134(29), 11974–11976. <https://doi.org/10.1021/ja305734x>
- Bludau, I., & Aebersold, R. (2020). Proteomic and interactomic insights into the molecular basis of cell functional diversity. *Nature Reviews Molecular Cell Biology*, 21(6), 327–340. <https://doi.org/10.1038/s41580-020-0231-2>
- Boersema, P. J., Raijmakers, R., Lemeer, S., Mohammed, S., & Heck, A. J. R. (2009). Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. *Nature Protocols*, 4(4), 484–494. <https://doi.org/10.1038/nprot.2009.21>
- Boesl, U. (2017). Time-of-flight mass spectrometry: Introduction to the basics: TIME-OF-FLIGHT MASS SPECTROMETRY. *Mass Spectrometry Reviews*, 36(1), 86–109. <https://doi.org/10.1002/mas.21520>
- 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., Daniels, D. L., & Yeo, G. W. (2016). SONAR Discovers RNA-Binding Proteins from Analysis of Large-Scale Protein-Protein Interactomes. *Molecular Cell*, 64(2), 282–293. <https://doi.org/10.1016/j.molcel.2016.09.003>
- Butter, F., Scheibe, M., Mörl, M., & Mann, M. (2009). Unbiased RNA–protein interaction screen by quantitative proteomics. *Proceedings of the National Academy of Sciences*, 106(26), 10626–10631. <https://doi.org/10.1073/pnas.0812099106>
- Cary, G. A., Vinh, D. B. N., May, P., Kuestner, R., & Dudley, A. M. (2015). Proteomic Analysis of Dhh1 Complexes Reveals a Role for Hsp40 Chaperone Ydj1 in Yeast P-Body Assembly. *G3 Genes|Genomes|Genetics*, 5(11), 2497–2511. <https://doi.org/10.1534/g3.115.021444>
- Casas-Vila, N., Sayols, S., Pérez-Martínez, L., Scheibe, M., & Butter, F. (2020). The RNA fold interactome of evolutionary conserved RNA structures in *S. cerevisiae*. *Nature Communications*, 11(1), 2789. <https://doi.org/10.1038/s41467-020-16555-4>
- Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B. M., Strein, C., Davey, N. E., Humphreys, D. T., Preiss, T., Steinmetz, L. M., Krijgsveld, J., & Hentze, M. W. (2012). Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. *Cell*, 149(6), 1393–1406. <https://doi.org/10.1016/j.cell.2012.04.031>
- Castello, A., Horos, R., Strein, C., Fischer, B., Eichelbaum, K., Steinmetz, L. M., Krijgsveld, J., & Hentze, M. W. (2013). System-wide identification of RNA-binding proteins by interactome capture. *Nature Protocols*, 8(3), 491–500. <https://doi.org/10.1038/nprot.2013.020>
- Ceron-Noriega, A., Almeida, M. V., Levin, M., & Butter, F. (2023). Nematode gene annotation by machine-learning-assisted proteotranscriptomics enables proteome-wide evolutionary analysis. *Genome Research*, 33(1), 112–128. <https://doi.org/10.1101/gr.277070.122>
- Chan, F. L., Vinod, B., Novy, K., Schittenhelm, R. B., Huang, C., Udugama, M., Nunez-Iglesias, J., Lin, J. I., Hii, L., Chan, J., Pickett, H. A., Daly, R. J., & Wong, L. H. (2017). Aurora Kinase B, a novel regulator of TERF1 binding and telomeric integrity. *Nucleic Acids Research*, 45(21), 12340–12353. <https://doi.org/10.1093/nar/gkx904>

- Chapman, J. D., Goodlett, D. R., & Masselon, C. D. (2014). Multiplexed and data-independent tandem mass spectrometry for global proteome profiling. *Mass Spectrometry Reviews*, 33(6), 452–470. <https://doi.org/10.1002/mas.21400>
- Chich, J.-F., David, O., Villers, F., Schaeffer, B., Lutomski, D., & Huet, S. (2007). Statistics for proteomics: Experimental design and 2-DE differential analysis. *Journal of Chromatography B*, 849(1–2), 261–272. <https://doi.org/10.1016/j.jchromb.2006.09.033>
- Chu, H.-P., Cifuentes-Rojas, C., Kesner, B., Aeby, E., Lee, H., Wei, C., Oh, H. J., Boukhali, M., Haas, W., & Lee, J. T. (2017). TERRA RNA Antagonizes ATRX and Protects Telomeres. *Cell*, 170(1), 86–101.e16. <https://doi.org/10.1016/j.cell.2017.06.017>
- Chu, H.-P., Froberg, J. E., Kesner, B., Oh, H. J., Ji, F., Sadreyev, R., Pinter, S. F., & Lee, J. T. (2017). PAR-TERRA directs homologous sex chromosome pairing. *Nature Structural & Molecular Biology*, 24(8), 620–631. <https://doi.org/10.1038/nsmb.3432>
- Coller, J. M., Tucker, M., Sheth, U., Valencia-Sanchez, M. A., & Parker, R. (2001). The DEAD box helicase, Dhh1p, functions in mRNA decapping and interacts with both the decapping and deadenylase complexes. *RNA*, 7(12), 1717–1727. <https://doi.org/10.1017/S135583820101994X>
- Corley, M., Burns, M. C., & Yeo, G. W. (2020). How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms. *Molecular Cell*, 78(1), 9–29. <https://doi.org/10.1016/j.molcel.2020.03.011>
- Cox, J., Hein, M. Y., Lubner, C. A., Paron, I., Nagaraj, N., & Mann, M. (2014). Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Molecular & Cellular Proteomics*, 13(9), 2513–2526. <https://doi.org/10.1074/mcp.M113.031591>
- Cox, J., & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, 26(12), 1367–1372. <https://doi.org/10.1038/nbt.1511>
- Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., & Mann, M. (2011). Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *Journal of Proteome Research*, 10(4), 1794–1805. <https://doi.org/10.1021/pr101065j>
- Curtis, N. J., & Jeffery, C. J. (2021). The expanding world of metabolic enzymes moonlighting as RNA binding proteins. *Biochemical Society Transactions*, 49(3), 1099–1108. <https://doi.org/10.1042/BST20200664>
- Cusanelli, E., & Chartrand, P. (2015). Telomeric repeat-containing RNA TERRA: A noncoding RNA connecting telomere biology to genome integrity. *Frontiers in Genetics*, 6. <https://doi.org/10.3389/fgene.2015.00143>
- Cusanelli, E., Romero, C. A. P., & Chartrand, P. (2013). Telomeric Noncoding RNA TERRA Is Induced by Telomere Shortening to Nucleate Telomerase Molecules at Short Telomeres. *Molecular Cell*, 51(6), 780–791. <https://doi.org/10.1016/j.molcel.2013.08.029>
- de Godoy, L. M. F., Olsen, J. V., Cox, J., Nielsen, M. L., Hubner, N. C., Fröhlich, F., Walther, T. C., & Mann, M. (2008). Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature*, 455(7217), 1251–1254. <https://doi.org/10.1038/nature07341>

- de Hoffmann, E. (1996). Tandem mass spectrometry: A primer. *Journal of Mass Spectrometry*, 31(2), 129–137.
[https://doi.org/10.1002/\(SICI\)1096-9888\(199602\)31:2<129::AID-JMS305>3.0.CO;2-T](https://doi.org/10.1002/(SICI)1096-9888(199602)31:2<129::AID-JMS305>3.0.CO;2-T)
- De Silanes, I. L., Graña, O., De Bonis, M. L., Dominguez, O., Pisano, D. G., & Blasco, M. A. (2014). Identification of TERRA locus unveils a telomere protection role through association to nearly all chromosomes. *Nature Communications*, 5(1), 4723.
<https://doi.org/10.1038/ncomms5723>
- Dodds, J. N., & Baker, E. S. (2019). Ion Mobility Spectrometry: Fundamental Concepts, Instrumentation, Applications, and the Road Ahead. *Journal of the American Society for Mass Spectrometry*, 30(11), 2185–2195. <https://doi.org/10.1007/s13361-019-02288-2>
- Dreyfuss, G., Kim, V. N., & Kataoka, N. (2002). Messenger-RNA-binding proteins and the messages they carry. *Nature Reviews Molecular Cell Biology*, 3(3), 195–205.
<https://doi.org/10.1038/nrm760>
- Dunham, W. H., Mullin, M., & Gingras, A.-C. (2012). Affinity-purification coupled to mass spectrometry: Basic principles and strategies. *PROTEOMICS*, 12(10), 1576–1590.
<https://doi.org/10.1002/pmic.201100523>
- Dutertre, M., Lambert, S., Carreira, A., Amor-Gu eret, M., & Vagner, S. (2014). DNA damage: RNA-binding proteins protect from near and far. *Trends in Biochemical Sciences*, 39(3), 141–149. <https://doi.org/10.1016/j.tibs.2014.01.003>
- Elias, J. E., & Gygi, S. P. (2007). Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods*, 4(3), 207–214.
<https://doi.org/10.1038/nmeth1019>
- Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., & Whitehouse, C. M. (1989). Electrospray Ionization for Mass Spectrometry of Large Biomolecules. *Science*, 246(4926), 64–71.
<https://doi.org/10.1126/science.2675315>
- Fradera-Sola, A., Nischwitz, E., Bayer, M. E., Luck, K., & Butter, F. (2023). RNA-dependent interactome allows network-based assignment of RNA-binding protein function. *Nucleic Acids Research*, gkad245. <https://doi.org/10.1093/nar/gkad245>
- Galej, W. P., Oubridge, C., Newman, A. J., & Nagai, K. (2013). Crystal structure of Prp8 reveals active site cavity of the spliceosome. *Nature*, 493(7434), 638–643.
<https://doi.org/10.1038/nature11843>
- Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W., & Gygi, S. P. (2003). Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proceedings of the National Academy of Sciences*, 100(12), 6940–6945.
<https://doi.org/10.1073/pnas.0832254100>
- Glisovic, T., Bachorik, J. L., Yong, J., & Dreyfuss, G. (2008). RNA-binding proteins and post-transcriptional gene regulation. *FEBS Letters*, 582(14), 1977–1986.
<https://doi.org/10.1016/j.febslet.2008.03.004>
- Gonatopoulos-Pournatzis, T., & Cowling, V. H. (2014). Cap-binding complex (CBC). *Biochemical Journal*, 457(2), 231–242. <https://doi.org/10.1042/BJ20131214>
- Hasan, A., Cotobal, C., Duncan, C. D. S., & Mata, J. (2014). Systematic Analysis of the Role of RNA-Binding Proteins in the Regulation of RNA Stability. *PLoS Genetics*, 10(11), e1004684. <https://doi.org/10.1371/journal.pgen.1004684>

- Hentze, M. W., Castello, A., Schwarzl, T., & Preiss, T. (2018). A brave new world of RNA-binding proteins. *Nature Reviews Molecular Cell Biology*, 19(5), 327–341. <https://doi.org/10.1038/nrm.2017.130>
- Hentze, M. W., & Preiss, T. (2010). The REM phase of gene regulation. *Trends in Biochemical Sciences*, 35(8), 423–426. <https://doi.org/10.1016/j.tibs.2010.05.009>
- Hesselberth, J. R. (2013). Lives that introns lead after splicing: Intron fates after splicing. *Wiley Interdisciplinary Reviews: RNA*, 4(6), 677–691. <https://doi.org/10.1002/wrna.1187>
- Hirashima, K., & Seimiya, H. (2015). Telomeric repeat-containing RNA/G-quadruplex-forming sequences cause genome-wide alteration of gene expression in human cancer cells in vivo. *Nucleic Acids Research*, 43(4), 2022–2032. <https://doi.org/10.1093/nar/gkv063>
- Hogan, D. J., Riordan, D. P., Gerber, A. P., Herschlag, D., & Brown, P. O. (2008). Diverse RNA-Binding Proteins Interact with Functionally Related Sets of RNAs, Suggesting an Extensive Regulatory System. *PLoS Biology*, 6(10), e255. <https://doi.org/10.1371/journal.pbio.0060255>
- Hsieh, E. J., Bereman, M. S., Durand, S., Valaskovic, G. A., & MacCoss, M. J. (2013). Effects of Column and Gradient Lengths on Peak Capacity and Peptide Identification in Nanoflow LC-MS/MS of Complex Proteomic Samples. *Journal of the American Society for Mass Spectrometry*, 24(1), 148–153. <https://doi.org/10.1007/s13361-012-0508-6>
- Hsu, J.-L., Huang, S.-Y., Chow, N.-H., & Chen, S.-H. (2003). Stable-Isotope Dimethyl Labeling for Quantitative Proteomics. *Analytical Chemistry*, 75(24), 6843–6852. <https://doi.org/10.1021/ac0348625>
- Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., & Mann, M. (2005). Exponentially Modified Protein Abundance Index (emPAI) for Estimation of Absolute Protein Amount in Proteomics by the Number of Sequenced Peptides per Protein. *Molecular & Cellular Proteomics*, 4(9), 1265–1272. <https://doi.org/10.1074/mcp.M500061-MCP200>
- Jorgenson, J. W. (2010). Capillary Liquid Chromatography at Ultrahigh Pressures. *Annual Review of Analytical Chemistry*, 3(1), 129–150. <https://doi.org/10.1146/annurev.anchem.1.031207.113014>
- Kanehisa, M., & Goto, S. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research*, 4. <https://doi.org/10.1093/nar/28.1.27>
- Karas, M., & Krüger, R. (2003). Ion Formation in MALDI: The Cluster Ionization Mechanism. *Chemical Reviews*, 103(2), 427–440. <https://doi.org/10.1021/cr010376a>
- Karas, Michael., & Hillenkamp, Franz. (1988). Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical Chemistry*, 60(20), 2299–2301. <https://doi.org/10.1021/ac00171a028>
- Karpievitch, Y. V., Dabney, A. R., & Smith, R. D. (2012). Normalization and missing value imputation for label-free LC-MS analysis. *BMC Bioinformatics*, 13(S16), S5. <https://doi.org/10.1186/1471-2105-13-S16-S5>
- Katahira, J. (2012). mRNA export and the TREX complex. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1819(6), 507–513. <https://doi.org/10.1016/j.bbagr.2011.12.001>

- Keilhauer, E. C., Hein, M. Y., & Mann, M. (2015). Accurate Protein Complex Retrieval by Affinity Enrichment Mass Spectrometry (AE-MS) Rather than Affinity Purification Mass Spectrometry (AP-MS). *Molecular & Cellular Proteomics*, *14*(1), 120–135. <https://doi.org/10.1074/mcp.M114.041012>
- Kilchert, C., Sträßler, K., Kunetsky, V., & Änkö, M. (2020). From parts lists to functional significance—RNA–protein interactions in gene regulation. *WIREs RNA*, *11*(3). <https://doi.org/10.1002/wrna.1582>
- Kirkpatrick, D. S., Gerber, S. A., & Gygi, S. P. (2005). The absolute quantification strategy: A general procedure for the quantification of proteins and post-translational modifications. *Methods*, *35*(3), 265–273. <https://doi.org/10.1016/j.ymeth.2004.08.018>
- Klaric, J. A., Wüst, S., & Panier, S. (2021). New Faces of old Friends: Emerging new Roles of RNA-Binding Proteins in the DNA Double-Strand Break Response. *Frontiers in Molecular Biosciences*, *8*, 668821. <https://doi.org/10.3389/fmolb.2021.668821>
- Klass, D. M., Scheibe, M., Butter, F., Hogan, G. J., Mann, M., & Brown, P. O. (2013). Quantitative proteomic analysis reveals concurrent RNA–protein interactions and identifies new RNA-binding proteins in *Saccharomyces cerevisiae*. *Genome Research*, *23*(6), 1028–1038. <https://doi.org/10.1101/gr.153031.112>
- Köcher, T., Swart, R., & Mechtler, K. (2011). Ultra-High-Pressure RPLC Hyphenated to an LTQ-Orbitrap Velos Reveals a Linear Relation between Peak Capacity and Number of Identified Peptides. *Analytical Chemistry*, *83*(7), 2699–2704. <https://doi.org/10.1021/ac103243t>
- Kolakowski, B. M., & Mester, Z. (2007). Review of applications of high-field asymmetric waveform ion mobility spectrometry (FAIMS) and differential mobility spectrometry (DMS). *The Analyst*, *132*(9), 842. <https://doi.org/10.1039/b706039d>
- Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D., & Nesvizhskii, A. I. (2017). MSFragger: Ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. *Nature Methods*, *14*(5), 513–520. <https://doi.org/10.1038/nmeth.4256>
- Larson, J. D., & Hoskins, A. A. (2017). Dynamics and consequences of spliceosome E complex formation. *eLife*, *6*, e27592. <https://doi.org/10.7554/eLife.27592>
- Lazar, C., Gatto, L., Ferro, M., Bruley, C., & Burger, T. (2016). Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. *Journal of Proteome Research*, *15*(4), 1116–1125. <https://doi.org/10.1021/acs.jproteome.5b00981>
- Li, J., Van Vranken, J. G., Pontano Vaites, L., Schweppe, D. K., Huttlin, E. L., Etienne, C., Nandhikonda, P., Viner, R., Robitaille, A. M., Thompson, A. H., Kuhn, K., Pike, I., Bomgarden, R. D., Rogers, J. C., Gygi, S. P., & Paulo, J. A. (2020). TMTpro reagents: A set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples. *Nature Methods*, *17*(4), 399–404. <https://doi.org/10.1038/s41592-020-0781-4>
- Licatalosi, D. D., & Darnell, R. B. (2010). RNA processing and its regulation: Global insights into biological networks. *Nature Reviews Genetics*, *11*(1), 75–87. <https://doi.org/10.1038/nrg2673>

- Ludwig, C., Gillet, L., Rosenberger, G., Amon, S., Collins, B. C., & Aebersold, R. (2018). Data-independent acquisition-based SWATH - MS for quantitative proteomics: A tutorial. *Molecular Systems Biology*, 14(8). <https://doi.org/10.15252/msb.20178126>
- Lundberg, E., Fagerberg, L., Klevebring, D., Matic, I., Geiger, T., Cox, J., Älgenäs, C., Lundberg, J., Mann, M., & Uhlen, M. (2010). Defining the transcriptome and proteome in three functionally different human cell lines. *Molecular Systems Biology*, 6(1), 450. <https://doi.org/10.1038/msb.2010.106>
- Makarov, A. (2000). Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis. *Analytical Chemistry*, 72(6), 1156–1162. <https://doi.org/10.1021/ac991131p>
- Makarov, A., & Denisov, E. (2009). Dynamics of ions of intact proteins in the Orbitrap mass analyzer. *Journal of the American Society for Mass Spectrometry*, 20(8), 1486–1495. <https://doi.org/10.1016/j.jasms.2009.03.024>
- Makarov, A., Denisov, E., Kholomeev, A., Balschun, W., Lange, O., Strupat, K., & Horning, S. (2006). Performance Evaluation of a Hybrid Linear Ion Trap/Orbitrap Mass Spectrometer. *Analytical Chemistry*, 78(7), 2113–2120. <https://doi.org/10.1021/ac0518811>
- Mallm, J.-P., & Rippe, K. (2015). Aurora Kinase B Regulates Telomerase Activity via a Centromeric RNA in Stem Cells. *Cell Reports*, 11(10), 1667–1678. <https://doi.org/10.1016/j.celrep.2015.05.015>
- Mamyrin, B. A. (2001). Time-of-flight mass spectrometry (concepts, achievements, and prospects). *International Journal of Mass Spectrometry*, 206(3), 251–266. [https://doi.org/10.1016/S1387-3806\(00\)00392-4](https://doi.org/10.1016/S1387-3806(00)00392-4)
- Marguerat, S., Schmidt, A., Codlin, S., Chen, W., Aebersold, R., & Bähler, J. (2012). Quantitative Analysis of Fission Yeast Transcriptomes and Proteomes in Proliferating and Quiescent Cells. *Cell*, 151(3), 671–683. <https://doi.org/10.1016/j.cell.2012.09.019>
- Matia-González, A. M., Laing, E. E., & Gerber, A. P. (2015). Conserved mRNA-binding proteomes in eukaryotic organisms. *Nature Structural & Molecular Biology*, 22(12), 1027–1033. <https://doi.org/10.1038/nsmb.3128>
- Mayr, C. (2017). Regulation by 3'-Untranslated Regions. *Annual Review of Genetics*, 51(1), 171–194. <https://doi.org/10.1146/annurev-genet-120116-024704>
- Michelmann, K., Silveira, J. A., Ridgeway, M. E., & Park, M. A. (2015). Fundamentals of Trapped Ion Mobility Spectrometry. *Journal of the American Society for Mass Spectrometry*, 26(1), 14–24. <https://doi.org/10.1007/s13361-014-0999-4>
- Miller, P. E., & Denton, M. B. (1986). The quadrupole mass filter: Basic operating concepts. *Journal of Chemical Education*, 63(7), 617. <https://doi.org/10.1021/ed063p617>
- Milo, R. (2013). What is the total number of protein molecules per cell volume? A call to rethink some published values. *BioEssays*, 35(12), 1050–1055. <https://doi.org/10.1002/bies.201300066>
- Minajigi, A., Froberg, J. E., Wei, C., Sunwoo, H., Kesner, B., Colognori, D., Lessing, D., Payer, B., Boukhali, M., Haas, W., & Lee, J. T. (2015). A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation. *Science*, 349(6245), aab2276. <https://doi.org/10.1126/science.aab2276>

- Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., Finn, R. D., & Bateman, A. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Research*, *49*(D1), D412–D419. <https://doi.org/10.1093/nar/gkaa913>
- Mitchell, S. F., Jain, S., She, M., & Parker, R. (2013). Global analysis of yeast mRNPs. *Nature Structural & Molecular Biology*, *20*(1), 127–133. <https://doi.org/10.1038/nsmb.2468>
- Moravec, M., Wischniewski, H., Bah, A., Hu, Y., Liu, N., Lafranchi, L., King, M. C., & Azzalin, C. M. (2016). TERRA promotes telomerase-mediated telomere elongation in *Schizosaccharomyces pombe*. *EMBO Reports*, *17*(7), 999–1012. <https://doi.org/10.15252/embr.201541708>
- Mostovenko, E., Hassan, C., Rattke, J., Deelder, A. M., van Veelen, P. A., & Palmblad, M. (2013). Comparison of peptide and protein fractionation methods in proteomics. *EuPA Open Proteomics*, *1*, 30–37. <https://doi.org/10.1016/j.euprot.2013.09.001>
- Müller, J. B., Geyer, P. E., Colaço, A. R., Treit, P. V., Strauss, M. T., Oroshi, M., Doll, S., Virreira Winter, S., Bader, J. M., Köhler, N., Theis, F., Santos, A., & Mann, M. (2020). The proteome landscape of the kingdoms of life. *Nature*, *582*(7813), 592–596. <https://doi.org/10.1038/s41586-020-2402-x>
- Nesvizhskii, A. I. (2014). Proteogenomics: Concepts, applications and computational strategies. *Nature Methods*, *11*(11), 1114–1125. <https://doi.org/10.1038/nmeth.3144>
- Neve, J., Patel, R., Wang, Z., Louey, A., & Furger, A. M. (2017). Cleavage and polyadenylation: Ending the message expands gene regulation. *RNA Biology*, *14*(7), 865–890. <https://doi.org/10.1080/15476286.2017.1306171>
- Olsen, J. V., de Godoy, L. M. F., Li, G., Macek, B., Mortensen, P., Pesch, R., Makarov, A., Lange, O., Horning, S., & Mann, M. (2005). Parts per Million Mass Accuracy on an Orbitrap Mass Spectrometer via Lock Mass Injection into a C-trap. *Molecular & Cellular Proteomics*, *4*(12), 2010–2021. <https://doi.org/10.1074/mcp.T500030-MCP200>
- Olsen, J. V., Macek, B., Lange, O., Makarov, A., Horning, S., & Mann, M. (2007). Higher-energy C-trap dissociation for peptide modification analysis. *Nature Methods*, *4*(9), 709–712. <https://doi.org/10.1038/nmeth1060>
- Olsen, J. V., Ong, S.-E., & Mann, M. (2004). Trypsin Cleaves Exclusively C-terminal to Arginine and Lysine Residues. *Molecular & Cellular Proteomics*, *3*(6), 608–614. <https://doi.org/10.1074/mcp.T400003-MCP200>
- Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A., & Mann, M. (2002). Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a Simple and Accurate Approach to Expression Proteomics. *Molecular & Cellular Proteomics*, *1*(5), 376–386. <https://doi.org/10.1074/mcp.M200025-MCP200>
- Ong, S.-E., & Mann, M. (2005). Mass spectrometry-based proteomics turns quantitative. *Nature Chemical Biology*, *1*(5), 252–262. <https://doi.org/10.1038/nchembio736>
- Ong, S.-E., & Mann, M. (2006). A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nature Protocols*, *1*(6), 2650–2660. <https://doi.org/10.1038/nprot.2006.427>

- Öztürk, M., Freiwald, A., Cartano, J., Schmitt, R., Dejung, M., Luck, K., Al-Sady, B., Braun, S., Levin, M., & Butter, F. (2022). Proteome effects of genome-wide single gene perturbations. *Nature Communications*, *13*(1), 6153. <https://doi.org/10.1038/s41467-022-33814-8>
- Perkins, D. N., Pappin, D. J. C., Creasy, D. M., & Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, *20*(18), 3551–3567. [https://doi.org/10.1002/\(SICI\)1522-2683\(19991201\)20:18<3551::AID-ELPS3551>3.0.CO;2-2](https://doi.org/10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2)
- Petti, E., Buemi, V., Zappone, A., Schillaci, O., Broccia, P. V., Dinami, R., Matteoni, S., Benetti, R., & Schoeftner, S. (2019). SFPQ and NONO suppress RNA:DNA-hybrid-related telomere instability. *Nature Communications*, *10*(1), 1001. <https://doi.org/10.1038/s41467-019-08863-1>
- Plaschka, C., Newman, A. J., & Nagai, K. (2019). Structural Basis of Nuclear pre-mRNA Splicing: Lessons from Yeast. *Cold Spring Harbor Perspectives in Biology*, *11*(5), a032391. <https://doi.org/10.1101/cshperspect.a032391>
- Pu, S., Wong, J., Turner, B., Cho, E., & Wodak, S. J. (2009). Up-to-date catalogues of yeast protein complexes. *Nucleic Acids Research*, *37*(3), 825–831. <https://doi.org/10.1093/nar/gkn1005>
- Ramanathan, A., Robb, G. B., & Chan, S.-H. (2016). mRNA capping: Biological functions and applications. *Nucleic Acids Research*, *44*(16), 7511–7526. <https://doi.org/10.1093/nar/gkw551>
- Rao, B. S., & Parker, R. (2017). Numerous interactions act redundantly to assemble a tunable size of P bodies in *Saccharomyces cerevisiae*. *Proceedings of the National Academy of Sciences*, *114*(45). <https://doi.org/10.1073/pnas.1712396114>
- Rappsilber, J., Mann, M., & Ishihama, Y. (2007). Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nature Protocols*, *2*(8), 1896–1906. <https://doi.org/10.1038/nprot.2007.261>
- Rissland, O. S. (2017). The organization and regulation of mRNA –protein complexes. *WIREs RNA*, *8*(1). <https://doi.org/10.1002/wrna.1369>
- Roepstorff, P., & Fohlman, J. (1984). Letter to the editors. *Biological Mass Spectrometry*, *11*(11), 601–601. <https://doi.org/10.1002/bms.1200111109>
- Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A., & Pappin, D. J. (2004). Multiplexed Protein Quantitation in *Saccharomyces cerevisiae* Using Amine-reactive Isobaric Tagging Reagents. *Molecular & Cellular Proteomics*, *3*(12), 1154–1169. <https://doi.org/10.1074/mcp.M400129-MCP200>
- Rudenko, G., & Van Der Ploeg, L. H. (1989). Transcription of telomere repeats in protozoa. *The EMBO Journal*, *8*(9), 2633–2638. <https://doi.org/10.1002/j.1460-2075.1989.tb08403.x>
- Scheibe, M., Butter, F., Hafner, M., Tuschl, T., & Mann, M. (2012). Quantitative mass spectrometry and PAR-CLIP to identify RNA-protein interactions. *Nucleic Acids Research*, *40*(19), 9897–9902. <https://doi.org/10.1093/nar/gks746>

- Schneider, C., Agafonov, D. E., Schmitzová, J., Hartmuth, K., Fabrizio, P., & Lührmann, R. (2015). Dynamic Contacts of U2, RES, Cwc25, Prp8 and Prp45 Proteins with the Pre-mRNA Branch-Site and 3' Splice Site during Catalytic Activation and Step 1 Catalysis in Yeast Spliceosomes. *PLOS Genetics*, *11*(9), e1005539. <https://doi.org/10.1371/journal.pgen.1005539>
- Schoeftner, S., & Blasco, M. A. (2008). Developmentally regulated transcription of mammalian telomeres by DNA-dependent RNA polymerase II. *Nature Cell Biology*, *10*(2), 228–236. <https://doi.org/10.1038/ncb1685>
- Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., & Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature*, *473*(7347), 337–342. <https://doi.org/10.1038/nature10098>
- Scigelova, M., & Makarov, A. (2009). Advances in bioanalytical LC–MS using the Orbitrap™ mass analyzer. *Bioanalysis*, *1*(4), 741–754. <https://doi.org/10.4155/bio.09.65>
- Shaw, J. B., Li, W., Holden, D. D., Zhang, Y., Griep-Raming, J., Fellers, R. T., Early, B. P., Thomas, P. M., Kelleher, N. L., & Brodbelt, J. S. (2013). Complete Protein Characterization Using Top-Down Mass Spectrometry and Ultraviolet Photodissociation. *Journal of the American Chemical Society*, *135*(34), 12646–12651. <https://doi.org/10.1021/ja4029654>
- Shevchenko, A., Tomas, H., Havli, J., Olsen, J. V., & Mann, M. (2006). In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nature Protocols*, *1*(6), 2856–2860. <https://doi.org/10.1038/nprot.2006.468>
- Shinoda, K., Tomita, M., & Ishihama, Y. (2010). emPAI Calc—For the estimation of protein abundance from large-scale identification data by liquid chromatography-tandem mass spectrometry. *Bioinformatics*, *26*(4), 576–577. <https://doi.org/10.1093/bioinformatics/btp700>
- Sinha, A., & Mann, M. (2020). A beginner's guide to mass spectrometry-based proteomics. *The Biochemist*, *42*(5), 64–69. <https://doi.org/10.1042/BIO20200057>
- Smits, A. H., & Vermeulen, M. (2016). Characterizing Protein–Protein Interactions Using Mass Spectrometry: Challenges and Opportunities. *Trends in Biotechnology*, *34*(10), 825–834. <https://doi.org/10.1016/j.tibtech.2016.02.014>
- Steen, H., & Mann, M. (2004). The abc's (and xyz's) of peptide sequencing. *Nature Reviews Molecular Cell Biology*, *5*(9), 699–711. <https://doi.org/10.1038/nrm1468>
- Sweet, T., Kovalak, C., & Collier, J. (2012). The DEAD-Box Protein Dhh1 Promotes Decapping by Slowing Ribosome Movement. *PLoS Biology*, *10*(6), e1001342. <https://doi.org/10.1371/journal.pbio.1001342>
- Taylor, J. A., & Johnson, R. S. (1997). Sequence database searches viade novo peptide sequencing by tandem mass spectrometry. *Rapid Communications in Mass Spectrometry*, *11*(9), 1067–1075. [https://doi.org/10.1002/\(SICI\)1097-0231\(19970615\)11:9<1067::AID-RCM953>3.0.CO;2-L](https://doi.org/10.1002/(SICI)1097-0231(19970615)11:9<1067::AID-RCM953>3.0.CO;2-L)
- The Gene Ontology Consortium, Carbon, S., Douglass, E., Good, B. M., Unni, D. R., Harris, N. L., Mungall, C. J., Basu, S., Chisholm, R. L., Dodson, R. J., Hartline, E., Fey, P., Thomas, P. D., Albou, L.-P., Ebert, D., Kesling, M. J., Mi, H., Muruganujan, A., Huang, X., ... Elser, J. (2021). The Gene Ontology resource: Enriching a GOLD mine. *Nucleic Acids Research*, *49*(D1), D325–D334. <https://doi.org/10.1093/nar/gkaa1113>

- Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., & Hamon, C. (2003). Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS. *Analytical Chemistry*, *75*(8), 1895–1904. <https://doi.org/10.1021/ac0262560>
- Tran, J. C., Zamdborg, L., Ahlf, D. R., Lee, J. E., Catherman, A. D., Durbin, K. R., Tipton, J. D., Vellaichamy, A., Kellie, J. F., Li, M., Wu, C., Sweet, S. M. M., Early, B. P., Siuti, N., LeDuc, R. D., Compton, P. D., Thomas, P. M., & Kelleher, N. L. (2011). Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature*, *480*(7376), 254–258. <https://doi.org/10.1038/nature10575>
- Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley Publishing Company. <https://books.google.de/books?id=UT9dAAAAIAAJ>
- Tutucci, E., & Stutz, F. (2011). Keeping mRNPs in check during assembly and nuclear export. *Nature Reviews Molecular Cell Biology*, *12*(6), 377–384. <https://doi.org/10.1038/nrm3119>
- Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., & Rothberg, J. M. (2000). A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature*, *403*(6770), 623–627. <https://doi.org/10.1038/35001009>
- Van Beneden, A., Arnoult, N., & Decottignies, A. (2013). Telomeric RNA Expression: Length Matters. *Frontiers in Oncology*, *3*. <https://doi.org/10.3389/fonc.2013.00178>
- Vermeulen, M., Hubner, N. C., & Mann, M. (2008). High confidence determination of specific protein–protein interactions using quantitative mass spectrometry. *Current Opinion in Biotechnology*, *19*(4), 331–337. <https://doi.org/10.1016/j.copbio.2008.06.001>
- Viceconte, N., Lorient, A., Lona Abreu, P., Scheibe, M., Fradera Sola, A., Butter, F., De Smet, C., Azzalin, C. M., Arnoult, N., & Decottignies, A. (2021). PAR-TERRA is the main contributor to telomeric repeat-containing RNA transcripts in normal and cancer mouse cells. *RNA*, *27*(1), 106–121. <https://doi.org/10.1261/rna.076281.120>
- Westmoreland, T. J., Olson, J. A., Saito, W. Y., Huper, G., Marks, J. R., & Bennett, C. B. (2003). Dhh1 regulates the G1/S-checkpoint following DNA damage or BRCA1 expression in yeast1. *Journal of Surgical Research*, *113*(1), 62–73. [https://doi.org/10.1016/S0022-4804\(03\)00155-0](https://doi.org/10.1016/S0022-4804(03)00155-0)
- Wilkinson, M. E., Charenton, C., & Nagai, K. (2020). RNA Splicing by the Spliceosome. *Annual Review of Biochemistry*, *89*(1), 359–388. <https://doi.org/10.1146/annurev-biochem-091719-064225>
- Yehezkel, S., Segev, Y., Viegas-Péquignot, E., Skorecki, K., & Selig, S. (2008). Hypomethylation of subtelomeric regions in ICF syndrome is associated with abnormally short telomeres and enhanced transcription from telomeric regions. *Human Molecular Genetics*, *17*(18), 2776–2789. <https://doi.org/10.1093/hmg/ddn177>
- Zhang, Z. (2004). Prediction of Low-Energy Collision-Induced Dissociation Spectra of Peptides. *Analytical Chemistry*, *76*(14), 3908–3922. <https://doi.org/10.1021/ac049951b>
- Zubarev, R. A., & Makarov, A. (2013). Orbitrap Mass Spectrometry. *Analytical Chemistry*, *85*(11), 5288–5296. <https://doi.org/10.1021/ac4001223>

Acknowledgements

This thesis pretty much summarises what has been a huge part of my life during the last five years. Through the process I have been accompanied and supported by many people.

Falk, thank you for mentoring and guiding me through this rocky path a PhD can be. In an alternative reality, I am still stuck with microscopy images and RNA-FISH protocols, so thank you for keeping a flexible mindset and understanding what I wanted to get from my PhD years. I did learn a lot, both technical and soft skills, throughout these years!

Miguel and **Vassilis**, thank you for giving me advice and meaningful insights during our Thesis Advice Committee meetings. Eventually, these meetings also included the collaboration of **Julian**; thank you as well for giving me a much needed RNA-biologist perspective on the project.

Katja, thank you for collaborating with me on Article II. I learned a lot getting to be a network-biologist with you for a couple months, and you made a scary and stressful process such as “PAPER REVISIONS” transit through a smooth and successful road.

Susanne, **Eva** and **Sabrina**, thank you for being part of my thesis committee defence.

Butter group, thank you for existing. Quite an easy task to be acknowledged for, right? But in essence that is what I am most grateful about. You existed, I got the chance to join you and you played a fundamental aspect in this play: providing a happy, supporting and healthy environment to thrive in. You, O.G. PhD students **Alina**, **Núria**, **Sabrina**, **Hanna**, **Lara**, **Merve** and **Katarina**, made me feel welcomed and included since the first day. **Alejo**, **Maya**, and **Michal**, my buddies in the office and on the omics section of the lab, thank you for all the tips and tricks shared during the years and the occasional “Could you please reboot my PC? - I am doing home office” situation. **Carisa**, you brought so much joy to the lab, and I am so grateful about that! Thank you for sharing such energy with me, it boosted my grey pulldown-days. To the new Butter generation, **Varvara**, **Rachel**, **Paddie**, **Sarah**, **Lars**, **Jan**, and **Alex**, thank you for keeping the spark on! Nothing like some enthusiastic new members to reignite that fire. Finally, to my one and only student, **Marie**, thank you for being so estelar. You did an awesome job!

Marion, thank you for your patience and perfecting my wet-lab skills. **Mario, Jasmine, Anja, Jimmy**, and **Amit**, thank you for keeping the proteomics core facility running. All the cool stories I have developed during my PhD are only possible because of your efforts on supporting proteomics research, from sample preparation and machine maintenance to scripts and code for data analysis. This acknowledgment is also extended to all other **IMB core facilities**, thank you for easing the research enterprise!

I am also grateful to the many friendly faces from my fellow IPP program buddies. **Alex, Eleftheria, Francesca, Gabrielle, Karla, Sabrina** and **Vanessa**, thank you for all the shared memories!

Vivien, we shared this experience together almost entirely, just one month apart! I have only good memories left from these years. Thank you for all the lab dances (no one understood better the joys of speedvac-side dancing!), after-work dinners, bike trips and countless other moments. See you in the future, friend!

Adrian, Daniel, Gabriel, Guillem, Marcos, Max i **Tanausú**, els meus bioquímics, kakis de la vida. Qui s'hauria imaginat, quan estudiàvem junts, que jo acabaria amb un doctorat al damunt? Estic segur que cap de nosaltres, poques indicacions en teníem. Sovint es diu que els amics de la universitat són efímers; estic realment agraït que no sigui el nostre cas. Ja fa més de 10 anys que ens coneixem i heu estat una constant font d'inspiració. Gràcies!

Aleix, Elisenda, Gemma, Hèlia, Laia, Maria del Mar i **Núria** aquesta tesi no seria possible sense vosaltres. Ho dic en un sentit metafòric, però en part, també, en un sentit bastant literal. Recordo perfectament una tarda a Palafrugell en què vosaltres, Aleix i Hèlia, em va ajudar a triar on anar a fer el doctorat. Aquesta anècdota exemplifica el que he rebut de tots vosaltres durant tants i tants anys: recolzament emocional, suport i amor. Gràcies, penya!

Papa, Mama i **Germana**. Al final som un producte del nostre entorn més directe i no n'hi podria haver cap de millor. He arribat fins aquí amb la vostra ajuda, gràcies per l'amor de tants anys!

Emily, my last acknowledgement words are for you. You are the take-home message from my PhD thesis, if you will. I could not be more grateful to have shared this whole experience with you. You made it funnier, easier and better. I can not wait to journey into our next adventure.




Albert Fradera Sola

Bioinformatician | Data Scientist | Proteomics & Genomics Scientist




I consider myself both collaborative and adaptable. I am always eager to take on new and diverse projects. I am very enthusiastic about sharing and continuously developing my data visualization skills as a means to enhance scientific communication. I am an open-minded person and enjoy fostering strong relationships with my colleagues.



Education

- Current | 2018
- **PhD Candidate in Quantitative Proteomics**
JGU University  Mainz, Germany
 - PhD thesis: Quantitative Mass Spectrometry-based Proteomics to Investigate RNA-protein Interactions
- 2017 | 2016
- **M.Sc. in Omics Data Analysis**
University of Vic  Barcelona, Spain
 - Master thesis: Soring function for finding *Lotium Perenne* DEGs
 - Erasmus + grant fellow
- 2016 | 2009
- **B.Sc. in Biochemistry**
University of Barcelona  Barcelona, Spain
 - Bachelor thesis: NMR studies on alpha-synuclein
 - Erasmus + grant fellow


Research Experience

- Current | 2023
- **Postdoctoral Researcher**
Friedrich Loeffler Institute  Greifswald, Germany
 - Postdoctoral researcher with Dr. Falk Butter (Falk.Butter@fli.de)
 - Publications: 1 first-author paper (in preparation), 1 co-author paper (in preparation)
- 2023 | 2018
- **PhD Candidate**
Institute of Molecular Biology  Mainz, Germany
 - PhD candidate with Dr. Falk Butter (Falk.Butter@fli.de)
 - Publications: 1 first-author paper, 6 co-author papers
- 2017
- **Master Student**
Institute of Biological, Environmental and Rural Sciences  Aberystwyth, UK
 - MSc student with Dr. Narcis Fernandez-Fuentes (naf4@aber.ac.uk)
 - Publications: 1 first-author paper, 1 co-author paper

Contact

 <https://afraderasola.github.io/>

 afraderasola@gmail.com

 +49 179 3469988

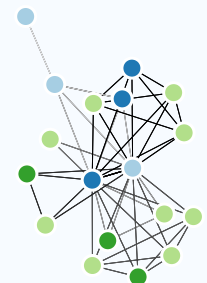
 [AFraderaSola](#)





 0000-0002-4780-9312

Language Skills



My CV as a network



Bullet points (for , , , and ) are interconnected by year.



Dry Lab Projects and Related Skills

• Network-based Assignment of RNA-Binding Proteins' Functionality

- Functionality was inferred from associations with annotations at the domain, molecular function, and pathway levels.
- Function-based networks were created from quantitative proteomics data and made accessible through an interactive, user-friendly web interface.
- **Keywords:** Network | Functional analysis | Pfam | GO | KEGG | R Shiny

• Infectome Profiling of Three *Leishmania* Species

- *M. musculus* proteome changes were investigated after infection with *Leishmania*.
- Protein orthologs between the *Leishmania* species were established.
- **Keywords:** Proteome | Differential expression | Protein orthology

• Embryonic Development Proteome Profiling of *Xenopus* Species

- Clustering (Self Organizing Maps) and dimensionality reduction techniques (Principal Component Analysis) were applied.
- **Keywords:** Proteome | Differential expression | SOM | PCA | Clustering

• Scoring Function for RNA-Seq Differential Expression Assessment

- A scoring function was developed around three RNA-Seq differential expression assessment R packages (DESeq2, edgeR, limma+voom)
- **Keywords:** NGS | RNA-Seq | Quality control | Differential expression

• ChIP-Seq Characterization of a Novel ATPase in *T. brucei*

- H2A.Z deposition was investigated and visualized with genome tracks
- **Keywords:** NGS | ChIP-Seq | Genome tracks



Wet Lab Projects and Related Skills

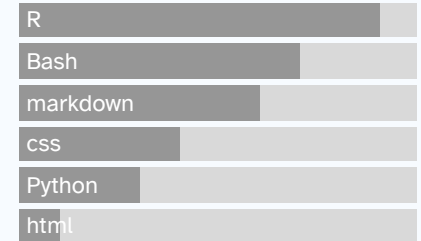
• Immunoprecipitation of RNA-binding Proteins in *S. cerevisiae*

- I was responsible for designing the experimental setup and generating the data for the RNA-binding protein interactome investigation.
- **Keywords:** Yeast | Protein Immunoprecipitation | Experimental design

• Mass Spectrometry Quantitative Proteomics

- I used chemical labeling (DML) and label-free mass spectrometry sample preparation protocols on several projects
- **Keyword:** Mass spectrometry | DML | LFQ | In-gel digestion

Coding Skills



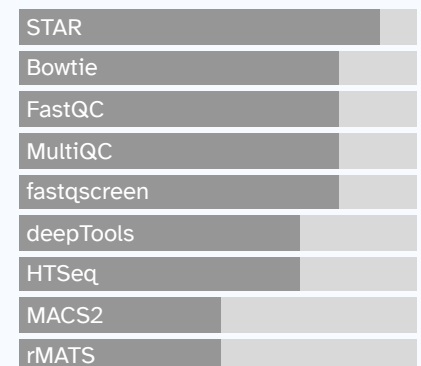
Some R packages I have experience with:



Proteomics Software Skills



Genomics Software Skills





Publications

Articles with Genomics Analysis

- 2022 | PLoS Pathogens | <https://doi.org/10.1371/journal.ppat.1010514>
- 2021 | Nature Communications | <https://doi.org/10.1038/s41467-021-22861-2>
- 2021 | PLoS One | <https://doi.org/10.1371/journal.pone.0249636>
- 2019 | PLoS One | <https://doi.org/10.1371/journal.pone.0220518>

Articles with Proteomics Analysis

- 2023 | Nucleic Acid Research | <https://doi.org/10.1093/nar/gkad245>
- 2023 | iScience | <https://doi.org/10.1016/j.isci.2023.106778>
- 2023 | PLoS Pathogens | <https://doi.org/10.1371/journal.ppat.1011486>
- 2021 | Journal of Cell Science | <https://doi.org/10.1242/jcs.254300>
- 2020 | RNA | <https://doi.org/10.1261/rna.076281.120>



Conferences and Courses

Conferences

- 2022 | Statistical data analysis for genome-scale biology | Flash talk
- 2021 | Network biology | Plenary talk
- 2019 | FEBS advanced course | Poster presentation

Courses

- 2022 | Data visualization for scientists
- 2021 | Scientific writing
- 2020 | Convincing scientific presentations
- 2020 | Regression Models | [Coursera certificate](#)
- 2020 | Statistical Inference | [Coursera certificate](#)

Abstract word cloud



Abstracts from the publications are summarized into a word cloud.

Document created with the R packages [pagedown](#) and [datadrivencv](#).

The source code is available on [github](#).

Last updated on 2023-10-27.