Applications of mass spectrometry-based proteomics: the developmental proteome of *D. melanogaster* and the RNA-fold interactome of conserved RNA structures in yeast

Dissertation zur Erlangung des Grades
Doktor der Naturwissenschaften

am Fachbereich Biologie
der Johannes Gutenberg-Universität in Mainz

Núria Casas-Vila, M.Sc.

Mainz, April 2020

Dekan:

Berichterstatter:

Berichterstatter:

Tag der mündlichen Prüfung:

Nothing in life is to be feared, it is only to be understood.
Now is the time to understand more, so that we fear less.

— Marie Curie

CONTENT

ABSTRACT

Mass spectrometry-based proteomics is a widely used technology that enables identification and quantification of proteins as well as their post-translational modifications. In this thesis, quantitative proteomics has been applied to the study of 1) protein developmental dynamics in *Drosophila melanogaster* and 2) the protein interactors to a set of evolutionary conserved RNA structures in yeast.

In **Chapter III** we apply label-free quantitative proteomics to comprehensively investigate protein remodelling across development in *Drosophila melanogaster*, a widely used model organism in genetic and developmental studies. To this end, we provide two datasets: the whole life cycle proteome consisting of 7952 proteins and the highly temporal-resolved embryogenesis proteome that comprises 5458 proteins. These large proteomic datasets allowed us to identify maternally provided proteins important in maternal-to-zygotic transition and early development, as well as stage- and gender-specific proteins. We also quantify isoform-specific expression of 34 different genes during development, validate expression of 268 small proteins and evidence the pseudogene Cyp9f3Ψ as a protein-coding gene. Integration with available transcriptomic data revealed moderate mRNA-protein correlation, with a protein delay of 4-5 hours. The combination of proteomic data with tissue-specific data uncovers proteins with tissue-specific developmental regulation, as exemplified by two yet uncharacterized proteins with an impaired muscle phenotype upon knockdown. The two large-scale proteomic datasets can be explored in detail in our interactive web interface and serve as a powerful resource for future studies.

Chapter IV focuses on the identification of RNA-binding proteins (RBP) associated with evolutionary conserved RNA structures in yeast. Using a SILAC-based quantitative RNA pull-down approach, we map 162 proteins to individual RNA structures within messenger RNA. The majority of them are associated with RNA-binding features, whereas approximately one third are

previously unrelated to RNA-binding and lack canonical RNA-binding domains. Intriguingly, we report a significant number of proteins binding to RNA folds in the coding regions of mRNAs, despite current knowledge about RNA-binding proteins regulation on the 5'- or 3'-UTR. Available PAR-CLIP datasets show high overlap with our reported protein-RNA interactions and RNA immunoprecipitation experiments confirmed our findings for selected mRNA-protein pairs. Using a reporter system and pulsed SILAC, we explored a role in translational control of a subset of mRNA-RBP pairs and find Nsr1 and YDR514C as possible translational regulators. This study presents a scalable immunoprecipitation-mass spectrometry (IP-MS) approach to map protein binders to individual RNA folds that connects structural RNA features to functionality.

Overall, during my PhD I applied quantitative MS-based proteomics to a broad range of biological questions, with special focus on the study of protein developmental dynamics and RNA-fold interactomics.

ZUSAMMENFASSUNG

Massenspektrometrie-basierte Proteomik ist eine weit verbreitete Technologie, die die Identifizierung und Quantifizierung von Proteinen sowie deren posttranslationale Modifikationen ermöglicht. In dieser Arbeit wurden mittels quantitativer Proteomik 1) die Proteindynamik während der Entwicklung von *Drosophila melanogaster* und 2) die Proteininteraktionspartner von evolutionär konservierten RNA-Strukturen in Hefe untersucht.

In **Kapitel III** nutzen wir quantitative Proteomik, um Änderungen der Proteinexpression während der Entwicklung von *Drosophila melanogaster*, einem in genetischen und entwicklungsbiologischen Studien weit verbreiteten Modellorganismus, umfassend zu untersuchen. Zu diesem Zweck erstellten wir zwei Datensätze: ein Lebenszyklus-Proteom mit 7952 Proteinen und ein Embryogenese-Proteom in hoher zeitlicher Auflösung mit 5458 Proteinen. Die großen Datensätze ermöglichten uns stadien- und geschlechtsspezifische Proteine sowie maternal bereitgestellte Proteine, die für den Übergang von Mutter zu Zygote und die frühe Entwicklung wichtig sind, zu identifizieren. Darüber hinaus haben wir eine isoformspezifische Expression von 34 Genen während der Entwicklung quantifiziert, die Expression von 268 kleinen Proteinen validiert und das Pseudogen Cyp9f3Ψ als proteinkodierendes Gen nachgewiesen. Die Integration von verfügbaren Transkriptom-Daten zeigte eine moderate mRNA-Protein-Korrelation mit einer Proteinverzögerung von 4 bis 5 Stunden. Die Kombination der Proteomdaten mit gewebespezifischen RNA-Expresssions daten deckte Proteine auf, die wichtig für die Entwicklung bestimmter Gewebe sind. Für zwei bisher noch nicht charakterisierte Proteine zeigten wir beispielsweise, dass ein Knockdown in einem beeinträchtigten Muskelphänotyp resultierte. Die umfassenden Proteom-Datensätze können über unsere interaktive Website genutzt werden und dienen als leistungsstarke Ressource für zukünftige Studien.

Kapitel IV befasst sich mit der Identifizierung von RNA-bindenden Proteinen, die mit evolutionär konservierten RNA-Strukturen in Hefe assoziiert sind. Wir

koppelten RNA-Pulldowns mit quantitativer Massenspektrometrie und ordneten 162 interagierende Proteine einzelnen RNA-Strukturen innerhalb der Messenger-RNA zu. Die Mehrheit der Proteine ist mit RNA-Bindungsmerkmalen assoziiert, während ungefähr ein Drittel keine kanonischen RNA-Bindungsdomänen aufweist und zuvor nicht mit RNA-Bindung in Zusammenhang gebracht wurde. Interessanterweise berichten wir über eine signifikante Anzahl von Proteinen, die an RNA-Faltungen in den kodierenden Regionen von mRNAs binden, obwohl sich das aktuelle Wissen über RNA-Bindungsproteine auf die 5'- oder 3'-UTRs konzentriert. Verfügbare PAR-CLIP-Datensätze zeigen eine hohe Überlappung mit unseren berichteten Protein-RNA-Interaktionen. Die Interaktion ausgewählter mRNA-Protein-Paare bestätigen wir zusätzlich durch RNA-Immunopräzipitation. Darüber hinaus untersuchen wir eine Untergruppe von mRNA-RBP-Paaren auf ihre Funktion in der Translationskontrolle mittels Reportersystem und pulsed-SILAC. Dadurch werden strukturelle RNA-Merkmale über die Interaktion mit Proteinpartnern mit ihrer Funktionalität verbunden.

Im Allgemeinen gilt dies für die Dauer der Behandlung und die Durchführung einer Reihe von Fragen, die sich insbesondere auf die Frage beziehen, ob es sich um eine besondere Frage handelt, die sich auf die Frage der Interaktion und der Interaktion mit der Arbeit bezieht.

STATEMENT OF CONTRIBUTION

The first authorship of the publication in **Chapter III** is shared between Alina, Sergi and myself. Alina and I planned, performed and quality-checked all sample collections at the university facilities of the Faculty of Genetics from JGU. I processed all collected time-points for MS analysis and supervised measurement together with Alina at the IMB. Mario performed first data analysis from raw files and data imputation. Sergi performed all data analysis included in the manuscript with input from with Alina, Falk and myself. Nadja performed *in vivo* experiments on selected candidates. During the course of the project, I discussed results and future directions together with Falk, Sergi, Alina and Mario in weekly meetings. I prepared all figures and wrote the article with Alina and Falk with input from Jean-Yves and Sergi. During revision, Alina and myself performed all experiments and made last changes on the manuscript.

In the project included in **Chapter IV**, I planned, executed and analyzed all experiments. Sergi performed the majority of data analysis and plots from different data sources. I wrote the manuscript and prepared all figures, with input from Falk. During revision, Lara and Marion helped with the last experiments.

Supervisor confirmation

LIST OF PUBLICATIONS

List of publications included in this thesis as separate chapters:

- The developmental proteome of *Drosophila melanogaster* (Chapter III).

- The RNA fold interactome of evolutionary conserved RNA structures in *S. cerevisiae* (Chapter IV).

Other publications during my PhD:

Becker, K., Bluhm, A., **Casas-Vila, N**., Dinges, N., Dejung, M., Sayols, S., Kreutz, C., Roignant, J. Y., Butter, F., & Legewie, S. (2018). Quantifying post-transcriptional regulation in the development of Drosophila melanogaster. *Nature Communications*, 9(1).

Winzi, M., **Vila, N. C.**, Paszkowski-Rogacz, M., Ding, L., Noack, S., Theis, M., Butter, F., & Buchholz, F. (2018). The long noncoding RNA lncR492 inhibits neural differentiation of murine embryonic stem cells. *PLoS ONE*, 13(1).

Bluhm, A., **Casas-Vila, N.**, Scheibe, M., & Butter, F. (2016). Reader interactome of epigenetic histone marks in birds. *Proteomics*, 16(3), 427–436.

Casas-Vila, N., Scheibe, M., Freiwald, A., Kappei, D., & Butter, F. (2015). Identification of TTAGGG-binding proteins in Neurospora crassa, a fungus with vertebrate-like telomere repeats. *BMC Genomics*, 16(1).

I. Introduction

INTRODUCTION

All cellular processes are dynamic and controlled. Proteins are key effectors of essentially all cellular processes. They carry out their functions at specific times and locations and in physical or functional association with other molecules. Protein abundance varies from a few copies to million copies per cell and the protein constituents of a cell are known as the proteome. Proteomes can dynamically adapt to internal or external stimuli and thereby determine the cellular phenotype. Gene expression regulation poses a challenge to proteomics since it increases protein diversity by alternative splicing. Post-translational modifications (PTMs) and polymorphisms are an additional source of protein variation. Moreover, protein-protein interactions and localization can influence protein function in the cell. Mass-spectrometry based quantitative proteomics enables the study of proteome structure and dynamics, which is central to understanding cellular functional states.

QUANTITATIVE MASS SPECTROMETRY-BASED PROTEOMICS

Mass spectrometry (MS)-based proteomics is an analytical technique that accurately measures the mass-to-charge ratio (m/z) of ionized molecules. In a typical MS workflow, entire proteins or peptides are separated by reverse-phase chromatography and are electrically charged by electrospray ionization (Fenn, Mann, Meng, Wong, & Whitehouse, 1989). Once ionized, peptides enter the vacuum of the mass spectrometer, and their m/z ratios are determined by their trajectories in the applied electromagnetic field (Walther & Mann, 2010) (Figure 1).

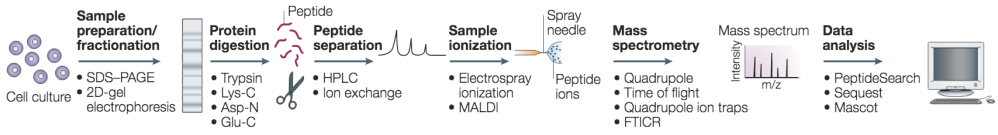


Figure 1. Common workflow in a bottom-up proteomics experiment. Proteins are extracted from the biological source, separated and digested into peptides. The resulting

peptides are separated via HPLC or ion exchange methods followed by ESI or MALDI ionization. Peptide ions enter the mass spectrometer and MS and MS/MS spectra are generated. Peptide sequence information is searched against protein databases for peptide identification and different software tools allow protein quantitation (Aebersold et al. 2016).

Mass spectrometry is not inherently quantitative since peptides exhibit different physicochemical properties that influence their ionization efficiency and hence their mass spectrometry detectability (Steen & Mann, 2004; Walther & Mann, 2010). To achieve accurate quantification, most proteomic workflows compare individual peptides between two different cellular states (e.g. treatment versus control) (Bantscheff et al., 2007; Bantscheff et al., 2012). Current quantification strategies can be classified in absolute and relative quantification methods. Absolute quantification infers the concentration of a protein by comparison to an added spiked-in standard with known concentration within the same sample. In contrast, relative quantification compares peptides in two different samples and calculates ratios to get a relative value for the overall protein (Lindemann et al., 2017). The rationale of the latter strategy relies on the fact that despite the peak size of a peptide is proportional to the number of peptide ions detected by the instrument, it is not possible to compare peaks of various peptides within a sample because of different ionization efficiencies. However, it is accurate to compare peaks of the same peptide – with the same ionization efficiency – in different samples.

Two distinct methodologies that differ in principle are generally used in proteomic studies: top-down and bottom-up approaches. Top-down proteomics analyses intact protein entities and allows precise identification of all protein products from a single gene (proteoforms). Because of high sequence coverage, top-down proteomics is beneficial to study splice variants, post-translational modifications (PTMs) and genetic variation (Tholey & Becker, 2017; Toby et al., 2016) and has been successfully applied in the field of degradomics – identification of protease substrate and products –, small proteins otherwise not detectable by peptide identification and for drug-target interactions (Durbin et al., 2016; Donnelly et al., 2019). Ionization and

fragmentation of whole proteins is usually achieved with electrospray ionization (ESI), generating multicharged protein ions and coupled to either Orbitrap mass analyzers or Fourier transform ion cyclotron resonance (FT-ICR) (Catherman, Skinner, & Kelleher, 2014). Despite software tools like ProSight are available, data interpretation is challenging since the number of fragment ions increases with protein size and might overlap, therefore complicating unambiguous assignment of fragment ions (Catherman et al., 2014; Tholey & Becker, 2017). In contrast, bottom-up proteomics achieves better quantification and is usually preferred for complex protein mixtures. Unlike top-down protocols, bottom-up workflows require endopeptidase digestion before analysis. The resulting peptides are separated by reverse-phase chromatography and coupled to electrospray ionization. In the vacuum of the mass spectrometer, peptide ions are fragmented in the gas phase to generate MS/MS (MS^2) spectra. MS^2 spectra information is collected and analyzed to accurately identify and quantify peptides by matching peptide masses to a protein database (Ruedi Aebersold & Mann, 2003; Steen & Mann, 2004). In this context, shotgun proteomics is referred to the use of bottom-up proteomics workflows to identify proteins from a complex sample mixture.

There is an intrinsic problem of mass spectrometers of not being able to fragment all precursor ions from MS^1 spectra because they are not fast enough to measure the entire precursor space at the MS^2 level. This challenge defines different modes of acquisition at the MS^1 level: data-dependent acquisition (DDA) and data-independent acquisition (DIA) (Ruedi Aebersold & Mann, 2016). DDA – also referred to as 'topN' – is widely used in discovery proteomics and isolates only the most intense ions for acquisition of MS^2 spectra. In contrast, DIA approaches continuously acquire MS^2 spectra of a selected mass window of precursor ions, generating series of very complex MS^2 spectra that are challenging to analyze. The Aebersold group introduced an approach termed SWATH-MS – sequential window acquisition of all theoretical fragment-ion spectra – that enables analysis of DIA data using prior knowledge of peptide fragmentation spectral libraries (Wolf-Yadlin et al., 2016; Ludwig et al., 2018;

Pappireddi et al., 2019). Unlike DIA and DDA, SRM/MRM is a hypothesis-driven strategy in which the known m/z of a peptide of interest is selected for fragmentation and monitored over time (Domon & Aebersold, 2006).

Data presented in this thesis were obtained following a bottom-up proteomics workflow using a DDA acquisition scheme with a Top10 selection method and different quantification strategies discussed in the following section.

QUANTIFICATION STRATEGIES

Data obtained both in the MS^1 level (peaks of precursor peptides) and the MS^2 level – fragment spectra information – can be used for quantification (Figure 2). Multiple methods offer high precision and accurate quantitation from MS^1 information. Probably the major limitation is that the MS^1 spectrum complexity increases with the number of samples to be compared that practically limits it to two or three samples. In contrast, up to 11-plex can be achieved with MS^2-based quantitation methods. However, these methods suffer from complex analysis of fragmentation peptides and poor quantification of low-abundant proteins (Pappireddi et al., 2019).

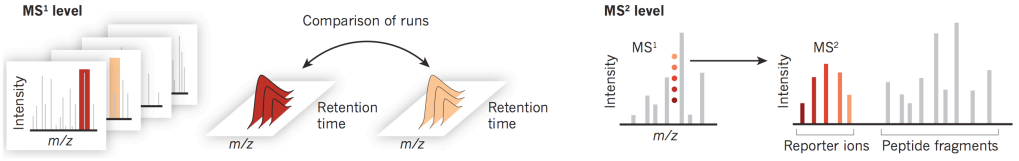


Figure 2. Quantification methods at the MS^1 and MS^2 level. Peptide quantification at the MS^1 level is based on the signal integration from peaks of the precursor ion from different aligned and normalized runs. MS^1 quantification strategies include comparison of isotopic peptides, absolute quantification and label-free quantification - when 'match between runs' option is activated, peptides identified from another run can be used to quantify a non-fragmented peptide if peak mass and elution time match -. In contrast, peptides can be quantified at the MS^2 level using fragment-ion intensities. This is the case of multiplexed shotgun proteomics, only after fragmentation they are distinguishable at the MS^2 spectra by distinct reporter ions.

Label-free quantification

Label-free quantification (LFQ) is simple and cost efficient, it does not require sample labeling and can therefore be applied to any kind of sample (Ong & Mann, 2005) (Figure 3). One form of LFQ quantification is based on counting the number of MS^2 spectra matching one protein – spectral counting – as a proxy for protein abundance. As a consequence, quantification of low abundant proteins is more variable and less reliable because a lower number of peptides are quantified (Bantscheff et al., 2012; Washburn et al., 2001; Lindemann et al., 2017). Alternatively, MS^1 level information can be used to estimate protein abundance based on the extracted ion current (XIC), which is proportional to peptide concentration (Pappireddi et al., 2019). In LFQ setups, multiple runs that need to be analyzed consecutively challenge experimental variability between samples. Also, a fraction of peptides will not be detected, a situation referred to the missing value problem that can be tackled by data imputation matching retention times and m/z values between samples (Washburn et al., 2001). Recently, Jürgen Cox proposed an algorithm for high accuracy quantification using XIC-based information, MaxLFQ, that is implemented within the MaxQuant software (Rgen Cox et al., 2014) –refer to *Data analysis* part for more information–. MaxLFQ enables robust and high accuracy quantification of large sample numbers.

In this thesis, the MaxLFQ algorithm was used in the study of *The Developmental proteome of Drosophila melanogaster* (Chapter III), in which 116 samples were analyzed –29 time-points in quadruplicates–. To maximize the number of peptide identifications, MaxLFQ was used in a "match between runs" mode that enables peptide quantification by matching MS^1 peptide identification across multiple samples, even though MS^2 data for a peptide is lacking.

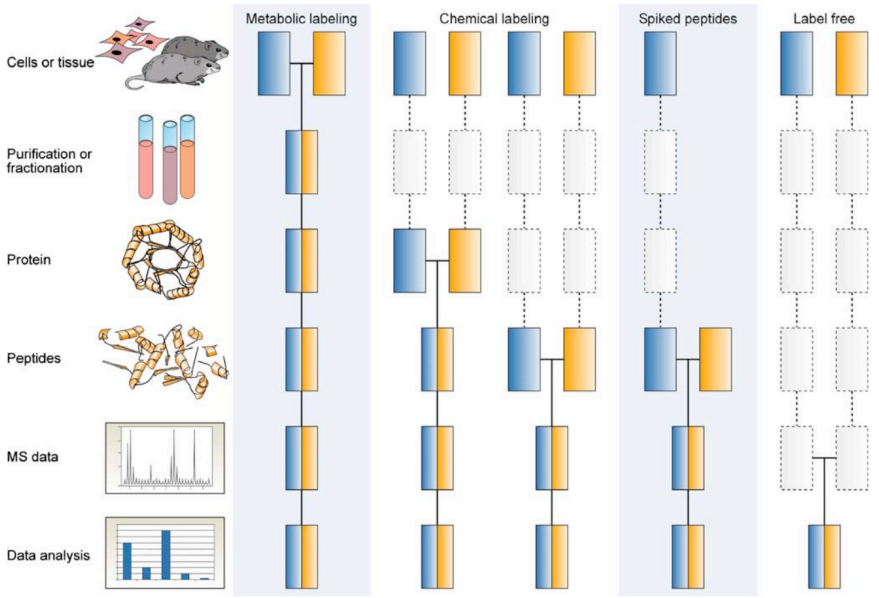


Figure 3. Scheme for common quantitative mass spectrometry strategies. Each box represents different experimental conditions, and dotted boxes indicate steps in which most experimental variation occur. Horizontal lines connecting blue and yellow boxes indicate points at which samples are combined (Bantscheff et al., 2007).

Heavy isotope labelling

These approaches are based on the principle that heavy isotopes have the same chemistry and elution patterns as their light counterparts and they are distinguishable by the mass spectrometer by a shift towards higher mass-to-charge ratios (Steen & Mann, 2004). Samples can be metabolically labelled at a cellular level by incorporation of heavy amino acids into the medium. Typically, one cell population is labelled with the 'light' version of an amino acid and another cell population is labelled with its 'heavy' counterpart. This leads to a known mass shift compared to the 'light' peptide, ideally higher than four Da (Bantscheff et al., 2007). After ensuring complete labelling, samples to be compared are combined before injection into the mass spectrometer, hereby relative quantification is achieved within the same experiment rather than between runs. This early sample combination represents a great advantage because it minimizes variations in the workflow due to separate handling and

offers high quantification accuracy and precision (Li et al., 2012) (Figure 3). First metabolic labelling approaches include the use of ^{15}N-enriched media in which all ^{14}N atoms are replaced by ^{15}N (Ong et al., 2002). This approach has been successfully applied in yeast, *Drosophila melanogaster* and other species. However, it does not ensure complete labelling and hence challenges data analysis (Conrads et al., 2001; Gouw et al., 2009). Alternatively, SILAC –stable isotope labelling by amino acids in cell culture– relies on the complete incorporation of heavy-labelled amino acids into newly synthetized proteins (Ong et al., 2002). Typically, heavy analogues of arginine and lysine contain ^{13}C and/or ^{15}N and at least one amino acid per peptide is labelled when proteins are trypsin-digested. Characteristic mass shifts of peptide pairs are detectable in the precursor spectrum and enrichment ratios between different experiments can be directly calculated. In addition to cultured cells, SILAC has been successfully applied to numerous organisms such as yeast (de Godoy et al., 2006) and whole animals like mouse (Krüger et al., 2008), *D. melanogaster* (Sury, Chen, & Selbach, 2010), *C. elegans* (Fredens et al., 2011) and zebrafish (Westman-Brinkmalm et al., 2011). SILAC can also be applied to non-dividing cells –e.g. whole tissue proteomes– as a spike-in standard to relatively quantify the non-labelled proteome compared to the SILAC protein mixture, a method termed super SILAC (Geiger et al., 2011). SILAC is considered one of the most reliable methods for protein quantification developed so far. Therefore the use of SILAC is widespread, ranging from protein turnover and translation studies – pulsed SILAC – (Schwanhäusser et al., 2009; Mathieson et al., 2018), protein methylation studies (Ong, Mittler, & Mann, 2004) and protein-protein or nucleic acid-protein interaction studies (Butter, Scheibe, Mö Rl, & Mann, 2009; Hubner et al., 2010; Mittler, Butter, & Mann, 2009). In the last years, SILAC advances that increase the multiplexing capabilities have been developed via the use of labels in smaller mass windows in the range of mDa (Hebert et al., 2013; Overmyer et al., 2018). In this thesis, SILAC was used in combination with RNA affinity purification in the study *The RNA fold interactome of evolutionary conserved RNA structures in S. cerevisiae* (Chapter IV).

Alternatively, peptides can be chemically modified by addition of a chemical tag. One example is isotope-coded affinity tag (ICAT) in which cysteine residues of peptides are covalently modified at the protein level and later affinity-purified (Gygi et al., 1999). This method ensures complete labelling of proteins; however, it is restricted to peptides containing cysteine (Pappireddi et al., 2019). Another widely used method is dimethyl labelling (DML). Here, primary amines are labelled through reductive amination by formaldehyde. In this reaction, isotopically labelled formaldehyde and sodium cynoborohydride ($NaBH_3CN$) converts primary amines found in the N-terminus and α-amino group of lysine residues to dimethylamines (Hsu & Chen, 2016). Therefore, all peptides are modified at least by the N-terminus and peptides with lysine residues contain multiple labels.

Generally, while MS^1-based isotope-labelling methods present high reproducibility, accuracy and measurement precision, the multiplexing capability is limited due to high complexity of the MS^1 spectrum.

Multiplexed proteomics: incorporation of isobaric tags

In contrast, isobaric labelling methods quantify different peptide populations on the MS^2 level. The major advantage is that MS^1 spectra complexity does not increase with multiplexing, because labelled peptides from different experiments appear at the same peak in the MS^1. Upon fragmentation, reporter ions from different samples can be distinguishable by mass changes in the MS^2. Two popular methods are iTRAQ –isobaric tag for relative and absolute quantification– and TMT –tandem mass tag– labelling that covalently introduce isobaric tags via amine-specific reactions (Stadlmeier, Bogena, Wallner, Wühr, & Carell, 2018; Wühr et al., 2012). They can be applied to any type of samples, iTRAQ allows multiplexing of up to 8 different conditions (Pierce et al., 2008) and TMT can be used for 11-plex samples (McAlister et al., 2012; Werner et al., 2012). Possible limitations of isobaric tagging are side reactions and co-eluting peptides that complicate data analysis (Thompson et al., 2003).

Absolute protein quantification

Absolute quantification allows determination of the exact protein concentration in a sample. Established methods like AQUA and QconCAT make use of isotope-labelled peptides of known concentration that are spiked into the sample and to which sample peptide intensities are compared (S. A. Gerber et al., 2003; Lindemann et al., 2017). AQUA strategy employs commercially synthetized proteotypic peptides that are quantified prior to addition to the sample (Kuster, Schirle, Mallick, & Aebersold, 2005). One limitation is that the isotope-labelled peptides are added late in the sample preparation workflow and therefore do not represent systematic errors from sample handling, unless sample loss in kept to a minimum. QconCAT addresses this challenge by using recombinant proteins with large numbers of concatenated peptides. The resulting QconCAT protein is expressed and isotope-labelled in *E. coli*. After purification, the protein is added to the sample and labelled peptides generated during protease digestion, thereby removing quantification biases due to systematic errors during biochemical workflows (Beynon, Doherty, Pratt, & Gaskell, 2005). Alternatively, SILAC labelled full-length proteins can be added to cell extracts before fractionation and used for absolute quantification. In this context, the Mann lab introduced PrEST –Protein Epitope Signature Tag– that employs a library of SILAC-labelled short and unique protein regions fused to purification and solubility tags for multiplexed quantification (Zeiler, Straube, Lundberg, Uhlen, & Mann, 2012). In contrast, iBAQ represents a label-free strategy for absolute protein quantification. iBAQ values are directly calculated by MaxQuant and represent de summed intensities of precursor peptides mapping to each protein divided by the number of theoretically tryptic peptides ranging from 6 to 30 amino acids in length (Krey et al., 2014).

THE MASS SPECTROMETER

The mass spectrometer consists of three main elements: the ion source, the mass analyzer and the ion detector. Numerous configurations exist for different

applications, each presenting distinct analytical performance. Some of these configurations will be discussed below, however, for all studies in this thesis a Q Exactive Plus instrument with a HPLC system coupled to ESI was used.

Peptide ionization

The ion source enables ionization of analyte molecules into the gas-phase. The development of so-called 'soft' ionization techniques such as electrospray ionization and matrix-assisted laser desorption/ionization (MALDI) represented a breakthrough in the proteomics field because they allowed transfer of highly polar, large molecules into the gas phase without fragmentation (Karas Michael, 1988). Generally, MALDI-MS setups are suitable for simple peptide mixtures and ESI is often the method of choice for complex protein mixtures. MALDI uses laser pulses on a dry matrix to sublime the non-volatile analyte molecules into the gas phase. Singly protonated ions formed by collision are then accelerated into the mass analyzer by electric potential (Steen & Mann, 2004). Alternatively, ESI volatizes and ionizes molecules out of a solution and it can be readily coupled to liquid chromatography. In ESI setups, peptides eluting from an HPLC column are electrostatically dispersed with high electrical potential held between the metal needle and the inlet of the mass spectrometer (Figure 4). As a result, large positively charged droplets containing both solvent and analyte are created. Solvent evaporates from the droplet increasing its density charge. Due to repetitive droplet fission and application of high electrical fields, smaller droplets are generated that enter the vacuum of the mass spectrometer for mass analysis.

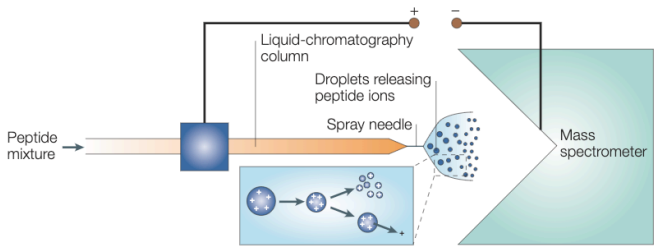


Figure 4. Electrospray ionization in LC-MS/MS. Peptides eluting from a chromatography column are electrostatically dispersed due to high electrical potential held between the tip of the metal needle and the entrance of the mass spectrometer. The resulting highly charged droplets are airborne and the solvent evaporates in repetitive fission cycles (Steen & Mann, 2004).

Mass analysis

The basic principle of a mass analyzer is separating ionized molecules according to their mass-to-charge ratio. The most common types include time-of-flight (TOF), quadrupole, ion trap, orbitrap and Fourier transform ion cyclotron (FT-ICR), and can be stand alone or in tandem configurations to exploit strengths of each (Ruedi Aebersold & Mann, 2003). TOF mass analyzers measure the m/z ratio of ions according to their flight time through a field-free tube under vacuum (R. Aebersold & Goodlett, 2001). Contrary, quadrupole and orbitrap mass analyzers use electrical/magnetic fields to accelerate and separate ions proportional to their charge-to-mass ratio. The quadrupole mass filter consists of four orthogonal metal rods connected in pairs. Opposing rods behave identically and apply constant (DC) or alternating (RF) voltages that affect ion trajectories through the length of the quadrupole. Specific voltage settings can be used to filter out highly charged ions or ions within specific mass ranges that will crash and not make it through the quadrupole. Thus, the quadrupole functions as a mass filter for ions within specific m/z isolation windows (Steen & Mann, 2004). The last generation mass spectrometers feature a segmented quadrupole that allows for narrower isolation windows and improves efficiency of low-mass isolation ranges (Domon & Aebersold, 2006). The orbitrap analyzer uses electrostatic fields to trap and isolate ions. It consists of a spindle-shaped electrode around which ion packets move in complex spiral patterns (Makarov, 2000). Ions not only oscillate in elliptical trajectories around the electrode but also move in axial trajectories that relate to their mass-to charge ratios. An outer electrode detects the image current of this axial motion and the signal is Fourier transformed to obtain high-resolution m/z values (Scigelova & Makarov, 2006).

Each type of mass analyzer presents distinct analytical capabilities based on its physical properties. For example, mass resolution is the minimum separation at which two equal peaks can be distinguished with a detectable 'valley' in between and mass accuracy is the difference between the measured and the theoretical mass in parts per million (ppm). These two concepts are intimately related because high-resolution leads to better mass accuracy (Xian, Hendrickson, & Marshall, 2012). Trap type mass analyzers like FT-ICR and orbitrap confer very high mass resolution (>100,000), very low accuracy at sub-ppm values and large dynamic range (Steen & Mann, 2004; Olsen et al., 2005). Other attributes such as the dynamic range –ability to detect low abundant ions–, scan speed –time to scan a certain m/z range– and peptide fragmentation efficiency are also important for the analytical performance. Beam type analyzers like quadrupole confer higher scan speed but lower mass resolution (Xian et al., 2012).

In this thesis, a hybrid instrument that combines the capabilities of two mass analyzers was used: a quadrupole analyzer for mass selection coupled to an orbitrap analyzer –see below the *Anatomy of a Q Exactive Plus* section–. This combination enables high speed, as well as very high accuracy and resolution (Michalski et al., 2011).

Ion fragmentation

In tandem MS, primary sequence information about selected peptides –product ions– can be obtained. This is achieved by coupling a first mass analyzer to a dissociation cell that allows fragmentation of selected precursor ions. The resulting product ions are analyzed in a second mass analyzer generating MS/MS spectrum (MS^2) (Figure 5a/5b).

Among common fragmentation methods we find electron transfer dissociation (ETD), electron capture dissociation (ECD) and higher-energy C-trap dissociation (HCD), each presenting different fragmentation features based on the underlying physical principle. While ETD is advantageous for larger and multi-charged peptide ions (Wiesner, Premsler, & Sickmann, 2008), ECD

preserves fragile posttranslational modifications and is often used in FT-MS (Wiesner et al., 2008; Zubarev, 2004). HCD is very effective for short and low charged peptide ions, enabling also good identification of modification sites (Sleno & Volmer, 2004). In our setup, precursor ions are selected by the quadrupole analyzer and fragmented in the HCD cell, where covalent bond breakage of the amide bonds is induced by collision with nitrogen gas (Shukla & Futrell, 2000). This creates complementary *b* and *y* ions, depending on whether the charge is retained in the C-terminus or N-terminus, respectively (Pappireddi et al., 2019). The resulting product ions are then analyzed in the orbitrap, from which the primary sequence of the precursor ion can be inferred (Figure 5c and 5d).

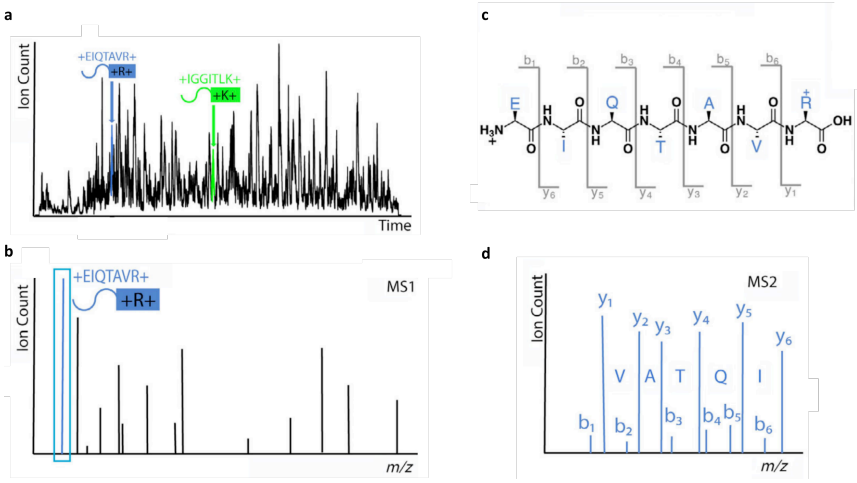


Figure 5. Outline of peptide identification and fragmentation. a LC-MS chromatogram shows all ion counts at each retention time. Two peptides (blue and green) elute at different times. **b** The MS¹ spectrum for the blue peptide is shown as a function of m/z values. Other co-eluting peptides can be distinguished by different m/z ratios. **c** chemical structure of a peptide with the designation of fragment ions after backbone fragmentation at the amide groups. Product ions are termed *b* when the charge is retained at the amino-terminal fragment and *y* when the carboxy-terminal part is charged. Ions are then labelled consecutively from the original terminus. **d** the MS² spectrum with fragment ions for the blue peptide is shown. The masses of the fragment ions together with the precursor mass from the MS¹ spectrum are used to unambiguously identify a peptide (adapted from Pappireddi N et al. 2019).

Anatomy of a Q Exactive Plus

The Q Exactive Plus features a quadrupole mass filter coupled to an orbitrap analyzer (Figure 6). It outperforms previous instruments in fragmentation speed thanks to parallel filling and detection. The instrument achieves high ion currents with S-lens and allows for cycle times of 1s for a top10 HCD method (Michalski et al., 2011).

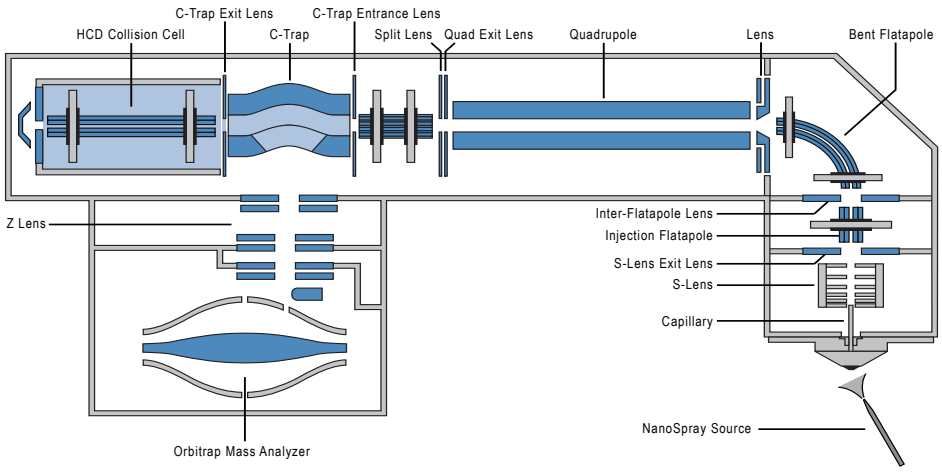


Figure 6. Schematic of the Q Exactive Plus from Thermo Scientific. First, ions produced at the source are captured and focused for effective transmission by different lenses and the bent flatapole filers out uncharged and neutral species. In a full MS^1 scan analysis, all ions are passed through the quadrupole and collected in the C-trap. After stabilization, they are transferred to the orbitrap for detection, where they trapped in orbital motion with their rotation frequency related to their mass-to-charge ratio. In contrast, in a MS/MS analysis, a specific set of ions is selected in the quadrupole and is transferred to the HCD collision cell for fragmentation. Similarly, the resulting product ions are then transferred back in the C-trap and send to the orbitrap for detection. Drawing not to scale (Michalski A et al. 2011).

SAMPLE PREPARATION AND CHROMATOGRAPHY-BASED PEPTIDE SEPARATION

Proper sample preparation is a critical step for accurate quantification of proteomes. Strategies differ depending on the biological source of the sample, complexity and localization of proteins. Also, depending on the analytical method, protocols can incorporate optimized lysis methods, subcellular

fractionation, depletion/enrichment of certain proteins or modifications and different mass tagging.

Cell lysis is the first step for efficient protein extraction and typically relies on physical and reagent-based methods. Physical lysis methods include dounce pestles, bead milling or sonication for cell disruption and are generally not appropriated for solubilizing membrane-associated proteins. Detergents are also used to disrupt cells and solubilize proteins. When downstream applications require preserving the native protein state, mild detergents such as Triton X-100 or NP-40 are commonly used. Alternatively, higher detergent concentrations or buffers with strong chaotropic agents like urea or guanidinium chloride result in protein denaturation. Because disruption of cellular compartments can influence protein stability by activating endogenous proteases and phosphatases, it is recommended to add protease and phosphatase inhibitors.

In complex samples high-abundant proteins can mask low-abundant protein detection. To prevent this, immunoaffinity columns can be used to remove high-abundant proteins or low-abundant proteins can be enriched according to their biochemical activity, cellular localization or post-translational modifications (Bellei et al., 2011). For example, density gradient centrifugation can be used to isolate specific subcellular fractions and phase-separating detergents to extract hydrophobic membrane components. Also, phosphorylated peptides can be enriched by IP using anti-phospho-specific antibodies, affinity ligands like ion-metal affinity chromatography (IMAC) or TiO_2 columns. Similarly, ubiquitinated peptides are typically enriched with diglycine antibody immunoprecipitation (Macek, Mann, & Olsen, 2009). An additional strategy to decrease sample complexity is SDS-PAGE gel fractionation, after which chopped gel pieces are processed for “in-gel” digestion (Shevchenko, Tomas, Havliš, Olsen, & Mann, 2007). Otherwise, protein digestion and subsequent steps are performed “in-solution”.

After protein extraction, a typical bottom-up MS workflow is followed by protein denaturation, reduction and alkylation. In in-gel digestion protocols

after SDS-PAGE gel separation, proteins are fixed with acetic acid and methanol and can be Coomassie-stained to visualize proteins before band excision or fractionation. After a destaining step with ammonium bicarbonate/ethanol and gel dehydration with acetonitrile, *in situ* disulfide bridge reduction with dithiothreitol (DTT) follows. Then sulfhydryl groups on cysteine residues are alkylated with iodoacetamide or iodoacetic acid to prevent disulfide bonds reforming. Proteins in the gel matrix can be fragmented into peptides by different endoproteinases such as trypsin, Lys-C, Glu-C and Asp-N, that diffuse into the gel and hydrolytically break peptide bonds (Shevchenko et al., 2007). The most commonly used protease is trypsin, which cleaves at the carboxyl-termini of lysine (K) and arginine (R). Peptides can be extracted from the gel with formic acid/acetonitrile. Alternatively, in-solution digestion protocols are detergent-free and instead use chaotropic agents like urea to denature proteins. The same protein reduction, alkylation and digestion steps follow (Havliš & Shevchenko, 2004). Generally, the amount of sample and its complexity are the main factors influencing the decision for in-gel or in-solution digestion protocols. In-gel digestion is good for complex sample separation and removal of detergents and salts; however low peptide recovery from the gel matrix results is significant peptide loss. In contrast, in-solution digestion is applicable to low amount samples, requires less processing time and has greater high-throughput potential.

An alternative method to obtain peptides from crude lysates is Filter-Aided Sample Preparation (FASP) (Wiśniewski, Zougman, Nagaraj, & Mann, 2009). FASP uses SDS to solubilize proteins, which is later dissociated in the presence of 8 M urea. Reduction and alkylation happen as usual and removal of low-molecular-weight components like detergents and DTT is achieved by ultrafiltration. Furthermore, the Mann laboratory introduced the in-StageTip method in which samples are processed from cell lysis to elution of purified peptides in a single StageTip (Kulak, Pichler, Paron, Nagaraj, & Mann, 2014). Also, Krijgsveld and colleagues proposed the single-pot solid-phase-enhanced sample (SP3) method that utilizes paramagnetic beads to capture, wash and

digest proteins on-beads (Hughes et al., 2019). When automated in a 96-well format for clinical samples, the SP3 workflow robustly quantifies 500-1,000 proteins from 100 to 1,000 cells (Müller et al., 2020).

A final desalting step is needed before peptide separation by LC to remove salts that might interfere with measurement. Desalting using StageTip with C18 material as solid phase is the most common strategy for peptide cleaning and concentration (Rappsilber, Mann, & Ishihama, 2007). In addition, interfering detergents can be removed with affinity columns or precipitating reagents.

Chromatography-based peptide separation

Good peptide separation is key to accomplish high-quality MS data. LC methods have been widely used to separate peptides prior to MS analysis (LC-MS) (Ruedi Aebersold & Mann, 2016). LC separates analytes on a mobile phase based on their affinity to a stationary phase. The development of high performance LC (HPLC) represented a significant improvement to the proteomics field, since a pressurized mobile phase could be passed through the column (Thakur et al., 2011). Our group uses reversed-phase HPLC columns packed with long hydrophobic alkyl chains (C18). Gradient elution of peptides based on their hydrophobicity can be reached by increasing amounts of acetonitrile with trifluoroacetic acid (TFA) (Rappsilber et al., 2007). Alternatively, because tryptic peptides have a net charge of minimum +2 in acidic pH, strong cation exchange (SCX) columns can be used to separate peptides according to pH. Column properties like length, diameter, elution gradient and packing material have a great impact on the resolving power – number of peaks separated from one another over total elution time– and influence the quality of the MS measurement (Fairchild, Walworth, Horváth, & Guiochon, 2010).

DATA ANALYSIS

Proteomic experiments generate complex data and therefore necessitate software tools that allow peptide/protein identification and quantification. Different software options exist and among the most used ones we find Skyline, Peaks, OpenMS and MaxQuant (Weisser et al., 2013). All data in this thesis were processed with MaxQuant that supports analysis of large-scale data obtained from a broad number of relative quantification techniques like SILAC, DML, TMT, iTRAQ and LFQ (Cox & Mann, 2008). MaxQuant incorporates its own peptide database search engine Andromeda for automated peptide identification (Cox et al., 2011a). Each MS^2 spectrum is searched against the database and additionally; MaxQuant uses a second peptide search algorithm that enables the identification of more than one peptide from each MS^2 spectrum. Thereby, co-fragmentation signals of additional precursors in complex samples are also used for identification and increase overall protein quantification (Cox et al., 2011b).

A degree of ambiguity exists on each peptide identification, therefore scoring filters that control for false-positive matches are applied. The Andromeda score evaluates how well an acquired fragmentation spectrum matches the theoretical spectrum. The posterior error probability (PEP) measures the probability of each peptide spectrum match (PSM) of being a false positive and integrates the Andromeda score for each PSM with multiple properties like peptide length, charge, missed cleavages and variable modifications. MaxQuant uses a target-decoy search strategy in which a second database is created by reversing all protein sequences (Elias & Gygi, 2007). At a peptide level, PEP scores obtained from a search in the reverse database are used to find the threshold controlling the FDR at 1%. The PSMs with PEP score over the threshold are accepted and hits in the reverse database are annotated with the "REV_" prefix in output tables.

Next, peptide identifications are assembled to obtain information at a protein level. Due to sequence redundancy, sometimes peptides cannot be uniquely

mapped to a single protein and MaxQuant reports protein identifications at a group level to avoid overcounting identifications (Tyanova, Temu, & Cox, 2016). Similarly, the protein group score reflects the probability of a true identification and is calculated based on the PEP score of peptides of a protein group. In the case that two different protein groups are identified by distinct peptides except for one, the two protein groups are not combined. If quantification is set to the option "unique peptides only" this shared peptide will not be used for quantification, in contrast to the "unique + razor peptides" quantification setting. Otherwise, razor peptides –peptides shared between different protein groups– are only used for quantification of the protein group with the larger number of identifications. Identical to peptide identifications, an additional FDR control is calculated at the protein level using protein group scores.

For downstream analysis, MaxQuant creates several output tables that report different values according to the quantification strategy. Some examples are protein intensities –sum of all identified peptide intensities for a protein group– , normalized protein ratios –e.g. SILAC ratios–, LFQ intensities –relative protein quantification values across all samples– (Cox et al., 2014) and iBAQ protein intensities based on absolute quantification (Schwanhüsser et al., 2011). First steps on downstream data analysis include filtering proteins identified based on single peptide matches, short peptides of less than seven amino acids and reverse-identified hits. Results can be visualized within the MaxQuant Viewer or output tables can be processed for customized analysis. In this thesis, normalized SILAC protein ratios and LFQ intensities were processed with customized R scripts.

In the first article of this thesis, the MaxLFQ algorithm implemented within MaxQuant was used for protein quantification (Cox et al., 2014). MaxLFQ achieves high precision and accuracy by comparing intensities of peptides across runs. First, retention times of all runs are calibrated to correct for systematic errors. The activation of "match between runs" option maximizes the number of peptides used for quantification. This way, if a peptide is measured in the MS^1 level but not selected for fragmentation, information from

another run can be used to identify that peptide. Peptide intensities are normalized using pairwise peptide ratios to account for sample variability and LFQ intensities for each protein are reported. MaxLFQ outperforms other methods like spectral counting or summed peptide intensities, and allows accurate quantification of three-fold changes in large sample numbers (Rgen Cox et al., 2014).

In the second article, we used SILAC-based quantification. SILAC achieves high precision by quantifying relative intensities of chemically equivalent peptides within the same run. The "requantify" option maximizes quantification of peptides with a single isotope pattern –either the light or heavy counterpart is missing–. Then, MaxQuant uses information from the MS^2 spectrum of the single peptide to search for the missing peak in the MS^1 plane. Protein ratios are calculated as the median of all SILAC peptide ratios and normalized to control for unequal protein amounts.

APPLICATIONS OF MS-BASED PROTEOMICS

Proteomics is a very versatile tool that enables a wide range of applications. Complete proteomes under normal and perturbed conditions –e.g. control versus treatment– have been obtained (De Godoy et al., 2008). Also, a recent study used fractionation gradients combined with MS-based proteomics to generate a comprehensive subcellular map of human proteins (Thul et al., 2017). Proteomics has been successfully applied to protein turnover studies using isotope labelling and measuring label fading over time as a measure of protein degradation (Visscher et al., 2016). Proteomics has also greatly increased the knowledge on cell signalling networks by measuring PTMs that result in a mass shift compared to the unmodified peptide counterpart. Large-scale PTM studies are challenging because they drastically increase the search space. Also, modified peptides are present in low amounts and they require specific enrichment protocols prior to MS measurement (Olsen & Mann, 2013;

Choudhary et al., 2009; Macek et al., 2009). MS-based proteomics has also been largely applied to the study of protein interactions with DNA, RNA, peptide or other proteins, a field referred to as interactomics.

Proteomics to measure expression dynamics

This field adds the time dimension to protein expression, reflecting the dynamic properties of cellular systems upon specific stimuli –e.g. drug treatment or developmental process–. It requires multiple MS analysis comparing proteomes sampled at several time-points. It has been applied to a broad range of studies like organelle protein dynamics following drug treatment (Rothstein et al., 2005), proteomes following developmental processes (Sun et al., 2015), turnover proteomics in response to external stimuli (Boisvert et al., 2012) and PTM levels upon cell differentiation (Ahmad & Lamond, 2014). Continuous advances in the field improved speed and resolution of instruments, thereby enabling high coverage measurements in a short time. For example, 4,399 yeast proteins –from 6,600 open reading frames– were quantified by SILAC in measurement runs of 2 hours (De Godoy et al., 2008). A general challenge in proteomics is the dynamic range, that limits the detection of lowly expressed proteins below the limit of detection. In other systems, complete coverage is also limited by the complexity of fragmentation spectra introduced by alternative splicing events, PTMs and polymorphisms (Mann, Kulak, Nagaraj, & Cox, 2013).

Interactomics

Proteins do not function in isolation but rather as part of complex interaction networks. MS-based proteomics can also be used to study protein-protein, protein-RNA and protein-DNA interactions. Proteins are obtained by affinity purification (AP) of tagged DNA, RNA, protein or peptide baits and coupled to mass-spectrometry measurement (AP-MS) (Dunham, Mullin, & Gingras, 2012). Antibodies targeting endogenous proteins are widely used for immobilization of the bait protein on a matrix. Also, antibodies targeting epitope-tagged

proteins like tandem affinity purification (TAP), FLAG, hemagglutinin (HA) and green fluorescence protein (GFP) tags. Affinity tags should be carefully selected so that it does not interfere with the function of the bait protein nor the expression or stability. Also, peptides carrying specific modifications, DNA and RNA baits can be biotinylated and immobilized on streptavidin-coated beads. Alternatively, structural tags with streptavidin affinity like S1 aptamers or MS2 aptamers can be used to study proteins bound to RNA baits (Oeffinger, 2012). In AP-MS it is advisable to mimic endogenous conditions in order to capture biologically relevant interactions. However, an intrinsic limitation of cell extract preparation is alteration of the natural macromolecular context. This can lead to aberrant interactions and non-specific background in the pull-down. Therefore, a basic principle in interaction quantitative proteomics is to use background proteins –proteins that similarly appear in the pull-down with the bait and control– to distinguish specific binders. Control experiments are often performed using IgG antibody, knockout or untagged version of the bait protein, unmodified peptide sequences or scrambled DNA/RNA sequences as baits. In combination with quantitative proteomics these strategies can be used to distinguish true interactors from background contaminants. Isotopic labelling approaches, like SILAC, or label-free quantification have been successfully applied in this field (Walther & Mann, 2010). Many large-scale interaction studies in human using antibodies; TAP-, GFP-, HA- or FLAG-tagged proteins have been performed. Also, yeast interaction studies revealed thousands of interactions using TAP-affinity purification approaches (Gingras, Gstaiger, Raught, & Aebersold, 2007; Krogan et al., 2006).

Interactions in the cell can be transient, low affinity and context-dependent. This challenges quantitation because only a subset of protein interactions will actually be detected by AP-MS. Chemical crosslinking reagents that form covalent bonds between protein-protein and/or protein-nucleic acid have been used to capture transient or labile interactions (Gingras et al., 2007). Also, a human study using BAC cell lines reported thousands of protein interactions *in vivo* and revealed that weak interaction dominate the protein network (Hein et

al., 2015). Alternatively, enzyme-mediated proximity labelling followed by affinity purification of the protein of interest enables identification of protein interactions *in vivo* (Roux, Kim, Raida, & Burke, 2012). For instance, BioID uses a biotin ligase (BirA) fused to the protein of interest that is able to biotinylate neighbouring proteins in excess of biotin. Recently, BioID combined with AP-MS provided a complete view on protein interaction networks at subcellular localization of human cells (X. Liu et al., 2018). Despite current limitations, recent technical advances in the interactomics field move towards a more comprehensive view of protein networks that can be applied to a wide range of cellular systems.

PERSPECTIVES IN MS-BASED PROTEOMICS

MS-based proteomics has enormously increased our knowledge on the dynamics, interactions, modifications and localization of proteins and definitely contributed to our understanding of their regulation at a systems level. Current instruments are not sensitive enough to cover the complete dynamic range, which limits the detection of lowly expressed proteins below the limit of detection. Current hybrid Orbitrap instruments cover a dynamic range of 5 orders of magnitude, as cellular protein concentrations can range over 12 orders of magnitude (Timp & Timp, 2020). Also, complete coverage is limited by the complexity of fragmentation spectra introduced by alternative splicing events, PTMs and polymorsphisms that greatly increase the peptide search space (Mann et al., 2013). In addition, complete multi-proteome comparison is currently a bottleneck in measuring time and is technically challenged by unstable sequential MS runs. Multiplexing has overcome this problem by enabling comparison of multiple conditions in the same MS run, however a maximum of 11-plex is currently supported for high accuracy quantification.

Single-cell transcriptomics enables measurement of transcripts on individual cells to profile their biological heterogeneity (Karaiskos et al., 2017). Current quantitative proteomics workflows allow for near-complete proteome measurement, however experiments are performed with thousands of cells and

require multiple replicates for reproducible quantitation. Very premature experiments on single-cell proteomics quantified most abundant proteins and mass cytometry experiments measured 100 different protein targets in single cells by antibody-based strategies (Doerr, 2019). Optimized sample preparation protocols that minimize peptide loss together with more sensitive instruments will be essential in this context. Recently, a microfluidic nanodroplet sample preparation approach coupled to ultrasensitive nanoLC-MS/MS was able to quantify 670 protein groups from an individual HeLa cell (Dou et al., 2019).

Beyond MS-based proteomics, alternative methods for protein sequencing can be envisioned. Fluorescent protein fingerprinting combines protein fragmentation and labelling of specific amino acids that are then imaged in subsequent cycles of Edman degradation (Swaminathan et al., 2018). Cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) uses oligonucleotide-labelled antibodies to profile transcriptomes and proteins in single-cells (Stoeckius et al., 2017). Technical advances in throughput capabilities and low-input samples will be key to the field and will need to be accompanied with bioinformatics tools that enable data analysis and interpretation.

DEVELOPMENTAL BIOLOGY OF DROSOPHILA MELANOGASTER

The fruit fly *Drosophila melanogaster* has been extensively used as a model organism in genetic and developmental studies. Despite having four pairs of chromosomes accounting for only around 5% of the human genome size, Drosophila presents a more compact genome than humans, with about 15,500 annotated genes (Adams et al., 2000). Because humans and flies have retained around 60% of the homologs from the common ancestor, a large number of disease-causing genes in humans have a counterpart in Drosophila, a feature that makes this organism especially suitable for studying human diseases.

Especially Drosophila has been widely used in the study of neurological diseases such as Alzheimer's, Parkinson's and Huntington's disease (Gerald M. Rubin et al., 2000). Other practical advantages that facilitate Drosophila studies are the handling ease, a short life cycle and large offspring. To date, key studies on neurological and rare diseases have been performed in the fruit fly and many genetic tools are readily applicable, that let researchers bypass ethical issues otherwise required for human studies (Bier, Harrison, O'connor-Giles, & Wildonger, 2018; Oriel & Lasko, 2018).

Drosophila life cycle

Throughout its life cycle, Drosophila undergoes complete metamorphosis that progresses across four different developmental stages: embryo, larva, pupa and adult fly. Female flies lay about 400 fertilized eggs during their lifetime that within the next 22-24 hours are able to develop from a single cell to an embryo. Embryos are ovoid and covered in a strong envelope –chorion–, from which two thin stalks project to prevent sinking in liquid medium. During the first minutes of embryogenesis, the zygote nucleus undergoes several mitotic divisions and nuclei migrate to the periphery of the egg –syncytial blastoderm stage–. Two hours later cellularization takes place, which transforms a multinucleate embryo into a multicellular stage –cellular blastoderm–. During these early repetitive mitotic cycles, the embryo uses maternally supplied RNAs and proteins (Wolpert, Tickle, & Martinez Arias, 2015). At two and a half hours of development, a major onset of zygotic transcription occurs and embryos begin to produce their own mRNAs and proteins in a process termed maternal-to-zygotic transition (DeRenzo & Seydoux, 2004). Maternally deposited transcripts and proteins are essential in early stage axis patterning in which the embryo is polarized by differential localization of maternal mRNAs in the egg (Tadros & Lipshitz, 2009). Next, embryonic development advances through other processes like gastrulation that will give rise to the different germ layers –endoderm, mesoderm and ectoderm–, body segmentation, head involution and organogenesis (Gilbert, 2000). All morphological changes during

embryogenesis are driven by tightly regulated mechanisms at a transcriptional, translational and epigenetic level in very narrow time windows. When fully developed, embryos hatch and the first instar larvae emerge. Larvae appear like white worm-shaped burrowers with black mouth able to move and feed. During larval development, larvae shed their cuticle twice –molting– to reach adult size and the imaginal discs cells continue to develop into adult body structures like head, legs, wings, thorax and genitalia. Each of the three larval stages last for about one day and when L3 stage is reached, larvae crawl out of their food to find a dry surrounding for pupation. Pupation is a static stage in which complete metamorphosis towards the adult fly occurs, body shortens and cuticle becomes hardened and pigmented. The ecdysone hormone secretion and its downstream transcriptional responses mainly mark this process (Kozlova & Thummel, 2002). Morphologically, cells from imaginal discs differentiate into wings, eyes, mouthparts, genitalia, thorax, abdomen and legs, and four days later adult flies emerge from the pupal case in a process termed eclosion (Celniker et al., 2009). For the first 4-8 hours, the body structures of young flies are not strong and mating is impeded; therefore virgin flies can be selected during this time window. Male and female flies are visually distinguishable by males being smaller and featuring a dark round abdomen and sex combs on their legs. When mating, female flies can store sperm and allow later internal fertilization (Thurmond et al., 2019). Females can lay dozens of fertilized eggs per day and adult flies live for up to eight weeks. However, the complete life cycle is temperature-dependent and lasts about 10 days at 25°C.

Genome-wide transcriptomic studies on Drosophila melanogaster

Since Drosophila genome sequence was published (Adams et al., 2000), numerous genetic studies investigated the functions of genes by phenotype association and molecular assays (Dietzl et al., 2007; St Johnston, 2002). Changes in gene expression following embryonic development have been reported from *in situ* hybridization studies. Tomancak and colleagues described

spatial and temporal expression patters of over 6,000 genes during embryogenesis (Tomancak et al., 2007) and Lécuyer and colleagues reported tight correlations between mRNA subcellular localization and protein function of 3,370 genes by high-resolution *in situ* hybridization (Lécuyer et al., 2007). While larger scale microarray data sets allowed quantification of expression patters during embryogenesis (Kalinka et al., 2010) and adult flies tissues (Chintapalli, Wang, & Dow, 2007), in 2011 Graveley and colleagues published the first complete developmental transcriptome combining tiling microarrays and RNA-seq data sets of 30 distinct developmental stages across Drosophila life cycle (Graveley et al., 2011). This work uncovered regulation of thousands of transcripts, splicing and RNA editing events and also annotated novel coding and non-coding genes. Later, Brown and colleagues addressed tissue-specific RNA expression in dissected organs and diverse cell lines (J. B. Brown et al., 2014). Complementary work on mRNA tissue-specific expression was done in Tomancak's laboratory using systematic imaging to asses subcellular distribution of about 6,000 mRNAs during oogenesis and provided a valuable resource for the community (Jambor, Surendranath, et al., 2014).

Studies on protein dynamics during development have been restricted to certain stages or tissues and use different quantification strategies that complicates data integration. Sowell and colleagues applied three different mass spectrometry workflows: online LC-MS/MS in a LCQ ion trap, strong-cation-exchange chromatography (SCX) followed by LC-MS/MS analysis and a home-built prototype gas-phase ion mobility spectrometry (IMS) method (Sury et al., 2010). Data were combined at a peptide level and proteins searched against NCBI database using the MASCOT search engine. As a result, they quantified 1699 proteins in nine time points of adult flies and reported differences in protein abundance between young and old adult flies. Application of quantitative proteomics approaches based on stable isotope labelling of peptides to whole organisms is challenging. Pioneer studies by Rudold Schoenheimer used ^{15}N to label *Caenorhabditis elegans* and *Drosophila* (Krijgsveld et al., 2003) and enabled the study of maternal-to-zygotic transition

(Gouw et al., 2009) and seminal fluid proteins (Findlay, Yi, MacCoss, & Swanson, 2008). However, ^{15}N labelling has some drawbacks because most peptides contain dozens of nitrogen atoms, therefore labelling is partial and peptide identification is demanding since mass shifts between labelled and unlabelled peptides are dependent on the number of labelled nitrogen atoms. Alternatively, SILAC was applied to Drosophila SL2 cell lines (Bonaldi et al., 2008) and to adult flies by feeding heavy lysine-labelled yeast to fly larvae and obtained labelled first filial generation (Sury et al., 2010). This study enabled proteome-wide comparison of adult male and female fly proteomes. Other studies also applied metabolic labelling to Drosophila in order to investigate larval and pupal stages (Chang et al., 2013) or embryos (Beati, Langlands, Have, & Müller, 2019; Gouw et al., 2009). Label-free proteomics have also been applied to Drosophila in numerous occasions to investigate oocyte maturation (Kronja et al., 2014), embryonic development (Fabre et al., 2016) and adult fly proteomes (Xing et al., 2014). However, previous work is restricted to specific stages or developmental processes and therefore has low overall proteome coverage limited by the time point or tissue of interest. Brunner and colleagues reported the largest protein catalogue in Drosophila, in which they identified a total of 9124 proteins, corresponding to 63% of the predicted proteome. Such high proteome coverage was achieved by sample diversity: two different Drosophila cell lines and four developmental stages. Additionally, several biochemical fractionation techniques were key to reduce sample complexity, including SDS-PAGE, hypotonic lysis, manual dissection, gradient centrifugation, gel filtration, ultracentrifugation, strong cation exchange chromatography and free flow electrophoresis (Bonaldi et al., 2008). Still, comprehensive studies on proteome remodelling across Drosophila development are challenged by its complicated life cycle and wide dynamic ranges.

POST-TRANSCRIPTIONAL GENE REGULATION BY RNA-BINDING PROTEINS

DNA is transcribed by RNA-Polymerase (Pol II) into messenger RNA (mRNA), which is co- and post-transcriptionally processed. Post-transcriptional gene regulation (PTGR) refers to the control of gene expression at the RNA level and is fundamental to maintain cellular homeostasis in different cell types and environments (Corbett, 2018). Essential players in this process are RNA-binding proteins (RBPs) that form nuclear ribonucleoprotein particles (RNPs) together with the nascent pre-mRNA and control the next steps of RNA processing. In the 5'end of the mRNA, the first nucleotide is modified by the addition of a 7-methylguanosine (m^7G) 'cap' structure. In higher eukaryotes, introns of pre-mRNAs are removed by co-transcriptional splicing and polyadenylation complexes are recruited to the 3'end of mRNAs when Pol II reads through the functional polyadenylation site (PAS). The mRNA export to the cytoplasm is also assisted by different sets of RBPs that determine its destiny: localization to specific organelles, degradation or cytoplasmic translation. For mRNAs directly coupled to cytoplasmic translation, components of the translation machinery scan the 5'end of mRNAs in search for an AUG start codon and subsequently the mRNA is translated into a protein (Carter et al., 2001; Hentze et al., 2018; Glisovic et al., 2008). In addition to RBPs, other regulatory mechanisms like microRNAs (miRNAs) and RNA modifications such as N^6-methyladenosine (m^6A) can also alter gene expression (Chekulaeva & Filipowicz, 2009; Zhao et al., 2016). Since the focus of this thesis is gene expression regulation by RNA-binding proteins, the following introduction describes key features of RNA molecules and RBPs that enable RNA-protein interactions in eukaryotes, with a special focus on *Saccharomyces cerevisiae*.

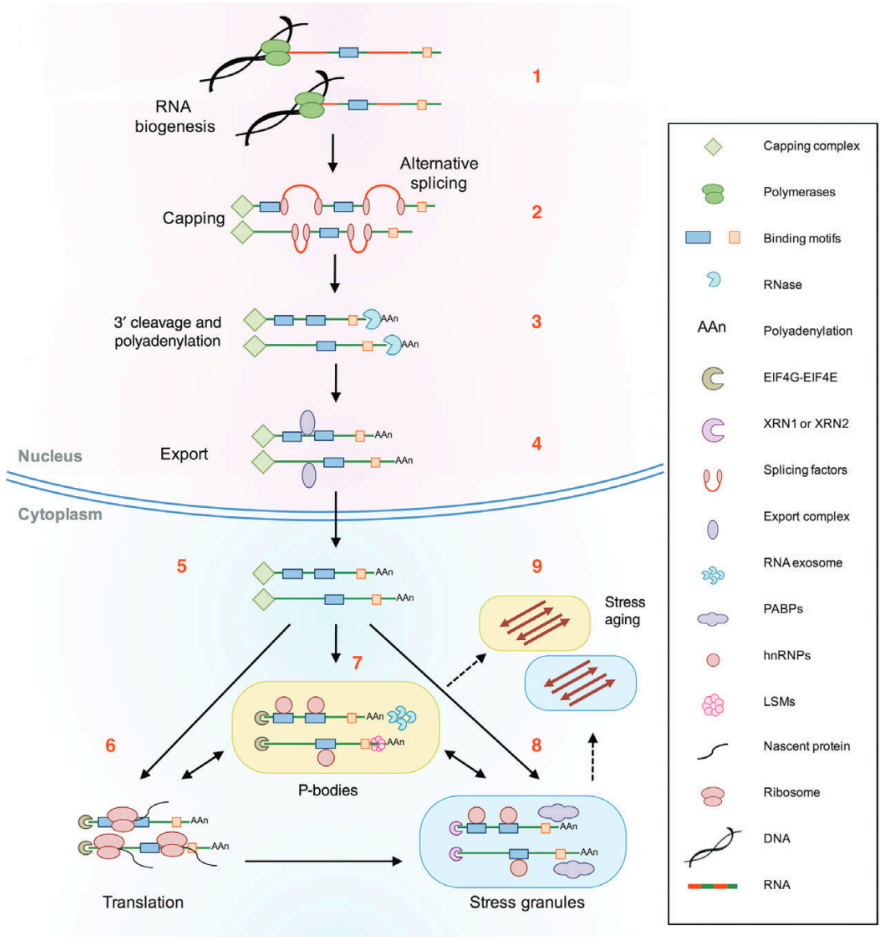


Figure 7. RNA-binding proteins influence every step of posttranscriptional regulation throughout mRNA lifetime. (1) RNA synthesis, (2) capping, splicing, and editing, (3) cleavage and polyadenylation, (4-5) nuclear export and cellular localization, (6) translation, (7) storage, (8-9) stability and degradation (Marchese, de Groot, Lorenzo Gotor, Livi, & Tartaglia, 2016).

STRUCTURAL INFORMATION ON RNA MOLECULES AND UNDERLYING FUNCTIONS

Beyond the information-encoding role of RNA, an additional layer of gene regulation is established by the propensity of RNA molecules to fold into three-dimensional (3D) structures. Already in prokaryotes we find prominent examples of RNA structure-directed functions. These include metabolite-

sensing RNAs (riboswitches) that can alter gene expression through direct changes in RNA conformation in response to cellular cues (Nahvi et al., 2002; Winkler et al., 2002) and thermosensors that take advantage of the temperature dependency of RNA folding to activate translation in *cis* (Klinkert & Narberhaus, 2009). In bacteria, the so-called ribozymes are RNAs with catalytic activities (Guerrier-Takada, Gardiner, Marsh, Pace, & Altman, 1963) able to cleave RNA without protein participation.

However, most structured RNA elements exert their function through interaction with RNA-binding proteins and/or *trans*-acting RNAs. To this end, RNA structures can be very diverse, ranging from base-pairing secondary structures in helical stems and hairpins, to long-range tertiary structures between distal RNA elements and even higher-order quaternary assemblies in complex with other molecules (Leontis et al., 2006; Ganser et al., 2019; Cruz & Westhof, 2009). An example of a functional stem-loop structure includes the internal ribosomal entry site (IRES) that is stabilized by interacting proteins and enables cap-independent translation initiation (S. A. Mitchell, Spriggs, Coldwell, Jackson, & Willis, 2003). More complex structures like the pseudoknot at the interferon gamma (*INFG*) mRNA can adjust its own expression to other immune genes in a feedback loop via the eIF2α (Cohen-Chalamish et al., 2009).

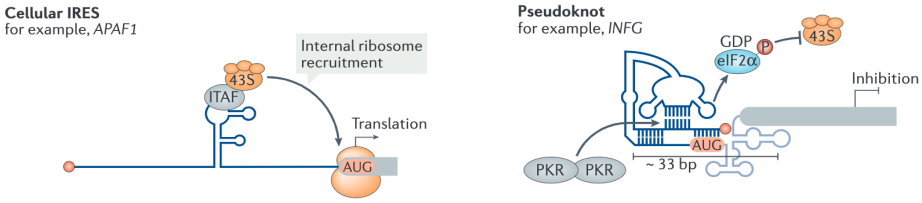


Figure 8. Examples of functional RNA structures. An IRES stem-loop structure in the apoptotic peptidase activating factor 1 (*APAF1)* mRNA can be stabilized by IRES *trans*-acting factors (ITAFs), thereby enabling the recruitment of the ribosome as an alternative to cap-dependent translation initiation. In contrast, the pseudoknot structure in the interferon gamma (*INFG*) mRNA adjusts its own expression to another member of the innate immunity pathway, the protein kinase R (*PKR*), via a feedback loop in which the activation of *PKR* results in phosphorylation of eIF2α and the consequent repression of *INFG* translation.

Not only the folding architecture has great impact on RNA functionality, but also its location within RNA molecules. UTRs are perceived as hotspots for RNA

structure-directed function. UTRs have greatly expanded during evolution, concomitant with higher eukaryotes presenting more complex regulatory mechanisms (Pesole et al., 2001). Also, interactions with regulatory factors is possible since they are usually not coated with translating ribosomes (Leppek, Das, & Barna, 2018). While mRNA structures positioned in 5' UTRs usually regulate translation in *cis* by blocking ribosome scanning or providing alternative IRESs, intronic structures contribute to splicing regulation and 3' UTR folds often influence mRNA decay and subcellular fates (Bartys et al., 2019; Pesole et al., 2001; Mayr, 2017). For instance, a stem-loop structure provokes exon skipping in the survival motor neuron 1 (*SMN1*) pre-mRNA and causes spinal muscular atrophy (Warf & Berglund, 2010; Bartys et al., 2019). Also, *oskar* mRNA is but one example of a plethora of localized mRNPs essential for early development in Drosophila. Its correct localization within the embryo is achieved by a stem-loop structure on the 3'UTR that enables dynein-dependent transport (Jambor, Mueller, Bullock, & Ephrussi, 2014).

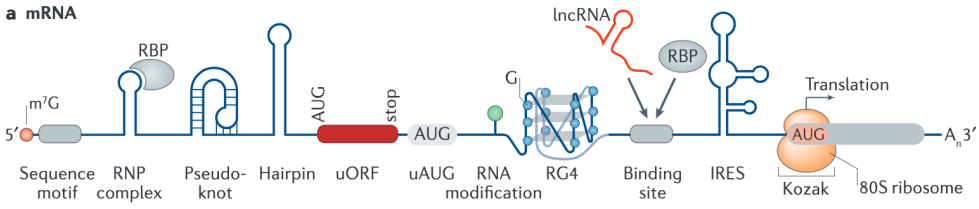


Figure 9. Overview of *cis*-acting regulatory elements and structures at the UTRs of mRNAs. 5' end 7-methylguanosine (m^7G) cap structures, sequence motifs, upstream open reading frames (uORF) and start codons (uAUG), RNA modifications and other structural elements like pseudoknots, hairpins, G-quadruplexes (RG4s) and IRES can influence translation. Structural elements at the 3' end as well as the poly(A) tail, stabilize the mRNA.

Another layer of complexity is added by the ability of RNA to dynamically shape its conformation in response to cellular cues, such as in the presence of RBPs, metabolites, ions, post-transcriptional modifications and temperature changes (Ganser et al., 2019). Because some conformations may form with high probability whereas some others may account for a small proportion of the RNA population, the distribution of all conformations of an RNA is referred to as an 'ensemble' (Ganser et al., 2019). It will be interesting to develop

experimental techniques that can tackle all plausible conformations of an RNA ensemble.

RNA structure profiling techniques

A wide variety of methods have been used to probe RNA structure. RNA foot printing has been extensively used for low-throughput RNA profiling. By treating for chemical modification or enzymatic digestion, unpaired nucleotides can be detected by reverse transcription termination in a polyacrylamide gel (Peattie & Gilbert, 1980). X-ray crystallography and NMR spectroscopy have also been used to determine high-resolution RNA structures within ribonucleoprotein complexes (Westhof, 2015). Advances over the last decade replaced capillary electrophoresis by deep sequencing and allowed high-throughput profiling to a whole transcriptome level (Kwok, Tang, Assmann, & Bevilacqua, 2015). The usage of chemicals that rapidly enter cells have also revolutionized the field and enabled *in vivo* RNA profiling (Huo et al., 2019; Kubota et al., 2015). To date, the most widely used chemical probes are dimethyl sulfate (DMS) and the conformational probing chemistry SHAPE – selective 2'-hydroxyl acylation with primer extension–. DMS alkylates accessible adenine and cytosine nucleotides (Cordero, Kladwang, Vanlang, & Das, 2012). Alternatively, SHAPE acylates the flexible 2'-hydroxyl (2'OH) common to all four nucleotides that results in reactivities correlating to Watson-Crick base pairing for every position of an RNA sequence (Merino et al., 2005; Wilkinson et al., 2006). In both strategies, modified RNA is coupled to deep sequencing after cDNA library preparation and structure probabilities are obtained by primer extension at a single-nucleotide resolution. Recently, DMS-seq was successfully applied to *S. cerevisiae*, *A. thaliana* seedlings and human cells and provided a comprehensive view of *in vivo* RNA structures as compared to *in vitro* conditions, highlighting the importance of cellular processes in shaping RNA structure (Ding et al., 2014; Rouskin et al., 2014; Wan et al., 2014).

In silico RNA folding predictions

Many computational efforts have focused on predicting RNA secondary structure based on free energy minimization algorithms (Rouskin et al., 2014). However, methods purely relying on sequence-based thermodynamics and compensating base pair changes, are able to accurately predict about 70% of RNA structure as quantified by *in vivo* experimental approaches (Mathews, 2004). This can be attributed to cellular factors influencing RNA structure inside the cells. Despite this, coupling these approaches with *in vitro* RNA structure data obtained by treatment with chemicals or nucleases have substantially improved the accuracy of these predictions (Mathews et al., 2004). Altogether, combining experimental approaches and modelling algorithms raises the prospect of identifying the complete folding state of all transcripts at a given cellular state and time.

RNA-BINDING PROTEINS

RNA-binding proteins bind mRNA throughout its lifetime and exert a tight control over all mRNA processing steps. They can act together with other RNAs and also work as connectors for other proteins that determine mRNA fate. To date, screens for proteins able to bind RNA have revealed a large number of RBPs in several organisms: of ≅20,500 protein-coding genes in humans, 7.5% are directly involved in RNA-binding (Scherrer, Mittal, & Janga, 2010; Tsvetanova et al., 2010; Baltz et al., 2012; Castello et al., 2012; Kwon et al., 2013; Hentze et al., 2018). Intriguingly, a shared observation is the detection of a wide range of proteins unrelated to RNA-binding that still remains elusive.

Functions of RNA-binding proteins

RBPs exert different functions on RNA molecules, ranging from mRNA processing, splicing, nuclear export, stability, folding, subcellular localization and translation. RBPs within mRNPs enable transport, control mRNA abundances and act as RNA chaperones to prevent mRNA misfolding. RBPs can

also regulate gene expression via interaction with long non-coding RNAs and piwi RNAs. Additionally, RBPs are constituents of the ribosomes and involved in the biogenesis of transfer RNAs (tRNAs), small nucleolar RNAs and ribosomal RNAs (rRNAs). Based on target-RNA classification, 50% of human RBPs function in mRNA metabolic pathways, whereas 11% are ribosomal proteins and the rest relate to other ncRNA metabolic processes (Gerstberger, Hafner, & Tuschl, 2014). Because RNP complexes assemble dynamically, the abundance of RBPs in the cell have a direct effect on RNA regulation. Also, competition among RBPs for overlapping targets can also define antagonistic or synergistic regulatory modes.

Dual functions of enzymes as RNA-binding proteins

Several studies identified a secondary role of metabolic enzymes as RBPs and their ability to regulate the expression of target mRNAs –moonlighting enzymes–, thereby providing regulatory links between metabolism and gene expression at certain cellular states (Castello, Hentze, & Preiss, 2015). For instance, phosphoglycerate kinsase (GAPDH) and thioredoxin (TYMS) represent novel members of the RBP class (Beckmann et al., 2015a). A prominent example is the Iron Regulatory Protein (IRP1)/Aconitase (ACO2) paradigm (Figure 11). ACO2 is a tricarboxylic acid (TCA) cycle enzyme that isomerizes citrate to isocitrate using a cubane iron sulfur cluster (4Fe-4S) as a cofactor (Hentze, Muckenthaler, Galy, & Camaschella, 2010). IRP1 shares 60% homology to ACO2, including conservation of the active site conferring IRP1 a dual function (Castello et al., 2015). Upon high iron levels, IRP1 functions as a cytosolic enzyme by binding the 4Fe-4S cluster. In iron-deficient cells, IRP1 disassembles the 4Fe-4S cluster and functions as RBP controlling iron homeostasis. To this end, IRP1 binds stem loop structures called iron responsive elements (IRE) in the UTRs of some mRNAs (Casey et al., 1988; Miillner & Kiihn, 1988). When iron is scarce, IRP1 binds a 5'UTR IRE of the iron storage protein ferritin and blocks its translation. Also, IRP1 stabilizes the

expression of the transferrin receptor by binding its 3'UTR IRE and thereby increasing iron uptake (Hentze et al., 2010).

In addition to enzymes, other proteins related to protein folding or actin-binding have also been identified as RBPs (Beckmann et al., 2015a). Nonetheless, the role of other protein classes in gene expression remains not clear and future studies will be important understand this regulation.

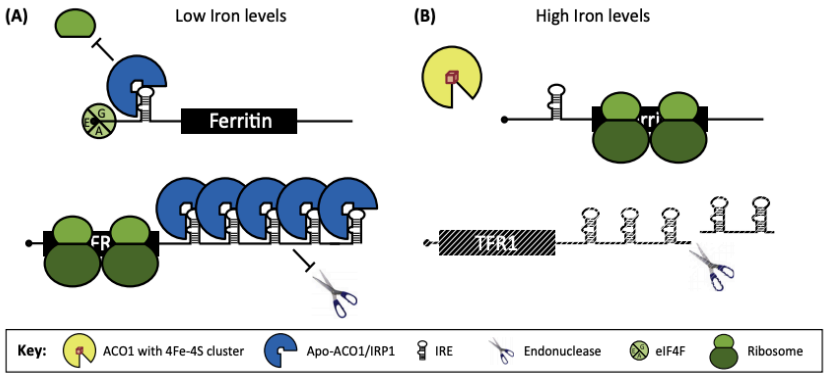


Figure 11. Dual role of the Iron Regulatory Protein 1 (IRP1) as a cytosolic enzyme and RNA-binding protein. Upon high iron levels, IRP1 functions as a cytosolic enzyme. In low iron conditions, IRP1 is not assembled with the 4Fe-4S cluster and binds 5' and 3' structure elements on iron metabolism genes, thereby inhibiting translation or stabilizing the target mRNAs.

Modular structure of RNA-binding domains

RNA binding domains enable proteins to contact RNA in a sequence- or structure-specific manner. The modes of action in mRNA fate control largely differ, ranging from binding to different *cis*-regulatory elements, collaborative interactions with other proteins in the same mRNA or antagonistic mechanisms competing for binding sites (Dassi, 2017). To date, a large set of RNA-binding domains has been proposed based on their occurrence in RNA-binding proteins. Among the most widespread domains we find RNA recognition motifs (RRM), K homology (KH) domains, C_3H Zinc finger (ZNFs), double-stranded RNA-binding motif (dsRBD) and the DEAD motif (Lunde, Moore, & Varani, 2007). Some other domains also include Piwi, Pumilio (PUF), Tudor and WD40 domains (Faoro & Ataide, 2014; Lunde et al., 2007).

Distinct protein domains employ different recognition modes on RNA. Within RRM, the β-sheet contacts four nucleotides of ssRNA through electrostatic interactions and hydrogen bonding. Additionally, RRMs can bind four to eight nucleotides via exposed loops and secondary structure elements of the peptide chains (Lunde et al., 2007). In contrast, C_3H Zinc finger domains contact RNA via hydrogen bonds between the protein backbone and the Watson-Crick edges of RNA bases. Instead, the side chains of the amino acids within the ZNF domain confer RNA-binding specificity by establishing contacts with the main chain and forming a structural template -hydrophobic binding pockets- for RNA recognition (R. S. Brown, 2005; Hall, 2005). As RBDs recognize few nucleotides or small structural motifs, a single RBD has little RNA-binding capability. Hence, many RBPs are arranged in a modular manner, consisting of RBDs and auxiliary domains in a single or tandem repeats. This increases binding diversity, affinity and specificity by contacting longer nucleotides sequences (Gerstberger et al., 2014). For example, PUF proteins use a repetitive and modular scaffold in which each repeat binds one nucleobase (Wang, Mclachlan, Zamore, & Tanaka Hall, 2002) and most ZNF proteins use tandem arrangements of ZNF or act as dimers to increase their RNA-binding affinity (R. S. Brown, 2005; Hall, 2005).

Transient and context-dependent interactions also challenge the detection of RBDs. In 2014, Gerstberger and colleagues established a census of human RBPs by database search and manual curation. They catalogued about 600 distinct RBDs linked to RNA-binding, from which 119 domains were exclusively associated with ribosomal proteins (Gerstberger et al., 2014). The resulting list includes domains found in proteins known to directly interact with RNA and domains associated with well-characterized mRNPs that might transiently contact RNA (Figure 10).

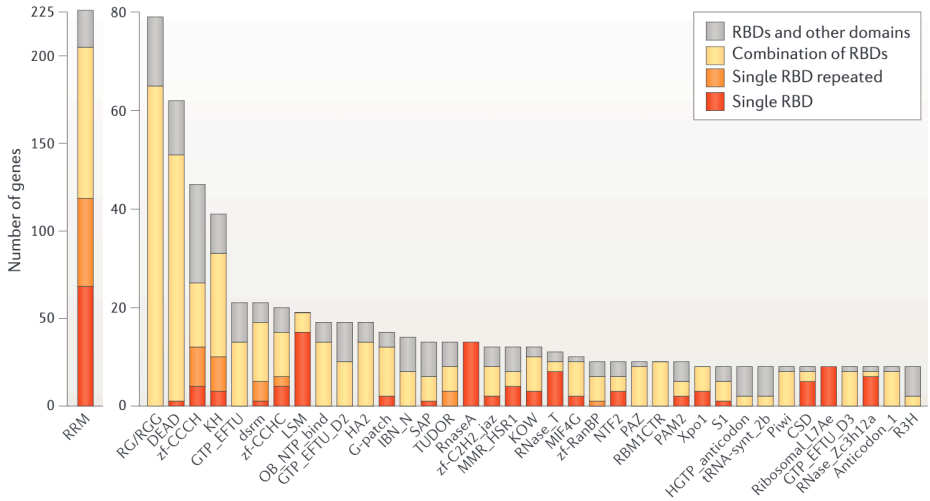


Figure 10. Most frequent human RBD and low-complexity RG/RGG-repeat regions (according to Pfam nomenclature). Proteins are classified according to the number of genes containing a single RBD (red), a combination of the same RBD (orange) or different RBD classes (yellow) and co-ocurrance with other domains unrelated to RNA function (grey) (Gerstberger et al., 2014).

Unstructured domains of RNA-binding proteins

In addition to the canonical RBD, linker regions between RBDs can form secondary structure elements that enlarge their RNA-binding surface and confer binding specificity. Intriguingly, recent methodological advances reported increasing amounts of RBPs lacking canonical RBDs. This can be explained because intrinsically disordered regions (IDRs) that lack 3D structures or unstructured low complexity domains are commonly present in RBPs (Phan et al., 2011). These regions are enriched for defined patterns of small, polar or charged amino acids like RG and YG and are often found in repeats (Beckmann et al., 2016; Järvelin et al., 2016). When hydrophobic peptides of the unstructured domain are exposed on the surface of a protein, they facilitate binding to other proteins or RNA (Habchi, Tompa, Longhi, & Uversky, 2014). Some of these unstructured domains are found in protein tails of ribosomal proteins and other RBPs, as well as in the disease-associated FMRP protein with an RGG box that recognizes guanidine-rich sequences (Klein et al., 2004; Thandapani et al., 2013; Phan et al., 2011).

METHODS TO STUDY RNA-PROTEIN INTERACTIONS

Because RNA-protein interactions are key to cellular homeostasis, an array of methods has been developed to study physical interactions between RNA and proteins. Since these interactions are often transient and their abundance vary depending on the subcellular location and environmental stimuli, different approaches can be selected according to the biological question. Such strategies are generally classified into RNA-centric and protein-centric methods. RNA-centric methods identify proteins associated to a specific RNA, whereas protein-centric approaches investigate RNAs that are bound by a protein of interest.

RNA-centric: identifying RBPs bound to a specific RNA

In vitro methods to study RNA-binding proteins include RNA pull-down experiments. *In vitro* transcribed RNAs are tagged and bound to a resin, incubated with cellular extract and unspecific binders washed away (Ramanathan et al., 2019; Marchese et al., 2016)(Figure 12). To this end, compounds like fluorescent dyes or digoxigenin can be used for RNA-tag-mediated purification (Faoro & Ataide, 2014). Another widely used strategy for RNA tagging is 5' or 3' biotinylation because of the high affinity of biotin to streptavidin (Lin, 2016). Alternatively, the minimized S1 aptamer sequence with high streptavidin affinity can be incorporated within RNA during *in vitro* synthesis (Srisawat et al., 2001; Srisawat & Engelke, 2001). All proteins can be eluted for analysis or biotin can be used to elute tagged RNA and its bound proteins if the tag is less affine than biotin (Leppek & Stoecklin, 2014; Klass et al., 2013). Elution of RNA and its associated proteins reduces background and excludes nonspecific interacting proteins bound to resin. A different approach that does not require the use of cellular extracts is hybridization of fluorescently labelled RNAs to a protein microarray (Kretz et al., 2013), but does not account for post-translational modifications of proteins and physiological concentrations. The ease and speed of *in vitro* methods enable the

possibility of mutagenesis studies that allow characterization of particular RNA-protein interactions –e.g. specific nucleotides or amino acids required for binding–. *In vitro* studies are also helpful to study proteins bound to low abundant RNA or low affinity and transient interactions that cannot otherwise be captured by *in vivo* cross-linking approaches. The use of a negative control RNA to assess binding specificity of positive results is recommended and it is often engineered by scrambling the sequence of the query RNA, in that way providing a same length and nucleotide composition control RNA, but with different primary and secondary structure.

In contrast, *in vivo* methods report physiologically relevant interactions. These include RNA or protein modifications, subcellular concentrations or localization relevant for binding. One popular technique is proximity proteomics that uses biotin ligases to covalently label proteins in close proximity (< 20 nm) and subsequent purification and identification by mass spectrometry (D. I. Kim et al., 2016). For example, the RNA-protein interaction detection method (RaPID) uses tagging of RNA with a BoxB aptamer that enables recruitment of the λ-N fusion protein and a promiscuous biotin ligase. In this way, lysine residues of nearby proteins –up to 66 nt away from the recruitment aptamer– get biotin-labelled (Ramanathan et al., 2018). While this approach does not require cross-linking or high cell numbers, it necessitates transfection of the RNA of interest. Moreover, this approach is more challenging for long RNAs as length and complex structures can limit accessibility of BoxB elements and consequently impede recruitment of biotin ligases. Other methods use protein-RNA crosslinking coupled to RNA purification under denaturing conditions to investigate proteins associated with a specific RNA *in vivo* (Figure 12). Distinct cross-linking conditions yield different efficiencies –e.g. formaldehyde, concentration, wavelength, energy and time– and therefore influence protein identification. Two different strategies are mostly used to cross-link protein and RNA. UV irradiation induces free radical formation at the nucleotide base that can attack amino acids in close proximity, thereby creating an irreversible covalent bond between protein and nucleic acid at zero distance (Ramanathan

et al., 2019; Ramanathan et al., 2018). UV does not capture protein-protein interactions because proteins do not absorb light at these wavelengths and therefore allows for the identification of direct RNA interactors. Alternatively, formaldehyde readily permeates cells and cross-links protein-protein, DNA-protein and RNA-protein molecules within 2 Å in a reversible manner. In contrast to UV light, formaldehyde cross-linking has also been used to study secondary protein-mediated interactions (Sugimoto et al., 2012) and has also been helpful to counteract UV light's low efficiency and biases towards certain base preferences or poorly cross-linked double-strand RNA regions (B. Kim & Kim, 2019; Ricci et al., 2014).

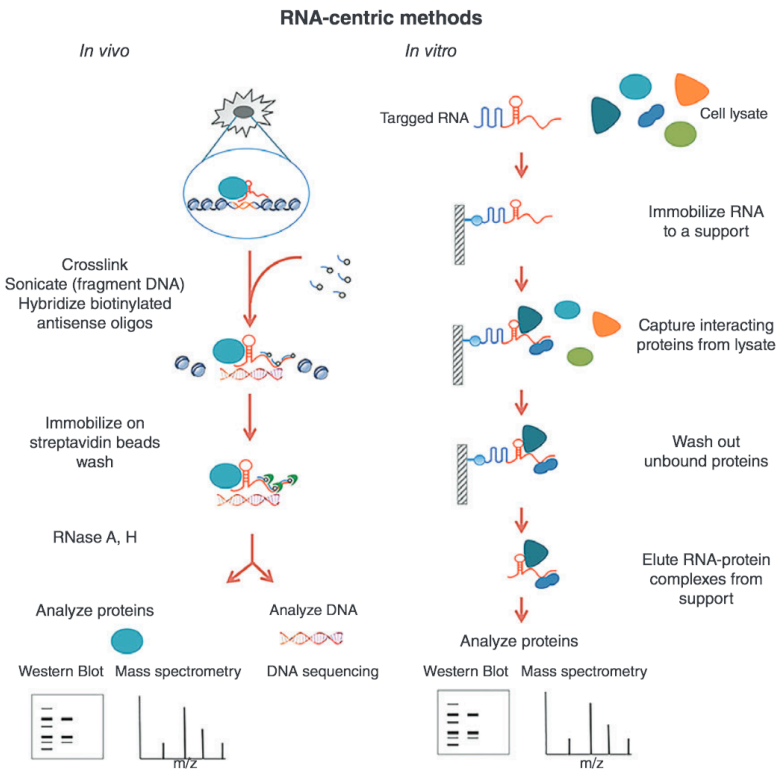


Figure 12. Schematics of RNA-centric methods to investigate RNA-binding proteins bound to an RNA of interest. *In vivo* approaches include cross-linking prior to cell lysis, affinity purification of protein-RNA-DNA complexes and a final RNA digestion step. *In vitro* strategies capture tagged or labeled RNA, unbound proteins are washed away and protein-RNA complexes are eluted from the immobilization matrix. Proteins bound to RNA can be either analyzed by western blot or mass spectrometry.

For example, chromatin isolation by RNA purification (ChIRP-MS) and capture hybridization analysis of RNA targets (CHART) use formaldehyde to cross-link RNA and proteins (Chu et al., 2015). Both methods capture cross-linked protein-RNA by hybridization with biotinylated oligonucleotide probes –with CHART requiring an additional RNase H digestion step to gain accessibility for the probes– (B. Kim & Kim, 2019; Ricci et al., 2014). In contrast, RNA affinity purification (RAP), peptide-nucleic-acid-assisted identification of RBPs (PAIR), MS2 *in vivo* biotin-tagged RAP (MS2-bioTRAP) and tandem RNA isolation procedure (TRIP) use UV cross-linking and different experimental setups (Ramanathan et al., 2019). Although these methods are conceptually the same because they all use oligonucleotide probes to hybridize UV cross-linked RNA followed by affinity purification, they incorporate some differences in their experimental workflows. For example, RAP uses long oligonucleotide probes and has been used to study protein interactors to long RNAs (Gaspar, 2018), whereas TRIP incorporates a previous poly(A) purification step (Matia-González, Iadevaia, & Gerber, 2017). PAIR employs oligonucleotide probes fused with cell penetrating peptides (Zeng et al., 2006). In the MS2-BioTRAP approach, the MS2 hairpin is fused to the RNA of interest and ectopically expressed together with the MS2 coat protein that is tethered to the MS2 hairpin and used to purify the RBP-bound RNA (Tsai, Wang, Huang, & Waterman, 2011). Generally, these methods require high input cell numbers and are coupled to proteomic analysis to discover the RBPs bound to RNA.

Protein-centric: discovering RNAs that bind a protein of interest

In vitro methods allow for characterization of physic-chemical properties of RNA-protein interactions. Among others, 'RNA-compete' uses incubation of RNA libraries with an immobilized protein of interest, followed by labelling of selected RNAs and hybridization to microarray detection (Ray et al., 2009)(Figure 13). In contrast, *in vivo* approaches capture a snapshot of RNAs interacting to a protein of interest at a certain cellular time and context. These methods use either direct purification of the protein of interest or rely on

selective modification of RNAs that interact with the selected protein. RNA immunoprecipitation (RIP) protocols capture RNAs bound to a protein of interest under native conditions (Keene, Komisarow, & Friedersdorf, 2006). UV light cross-linking at 254 nm covalently links RNA to almost all amino acids except from D, E N and Q (Kramer et al., 2014). This property gives rise to all the extensively used CLIP methods –cross-linking immunoprecipitation– that are coupled to high-throughput sequencing (HITS) (Sugimoto et al., 2012).

Each CLIP method (HITS-CLIP, PAR-CLIP, iCLIP, eCLIP, GoldCLIP, etc.) requires tailored downstream bioinformatics analysis due to differences in library preparation protocols (Ramanathan et al., 2019; Kishore et al., 2011). For example, both CLIP and iCLIP ligate the 3' adaptor, purify the cross-linked protein-RNA complex and include a protein digestion step. However, while CLIP captures cDNAs that read-through the cross-link site via 5' adaptor amplification, iCLIP captures cDNAs truncated at the cross-linking site by circularization and subsequent linearization, enabling nucleotide-resolution quantification (Sugimoto et al., 2012). To date, RNA target information on many RBPs have been produced using RBP-specific antibodies in different CLIP setups and are mostly available at http://www.endocodeproject.org. However, these methods are challenged by low recovery of cross-linked RNA due to poor cross-linking efficiency, low abundance of RNA-protein complexes and poor immunoprecipitation by antibodies. Alternatives like PAR-CLIP that uses 4-thiouridine and/or 5-thioguanine as nucleotide analogues can be helpful when UV cross-linking is not efficient (Hafner et al., 2010). Additionally, formaldehyde cross-linking for dsRNA-binding proteins or other emerging cross-linking compounds like DTT, diepoxybutane and 2-iminothiolane have been also used in some setups (B. Kim & Kim, 2019; Ramanathan et al., 2019).

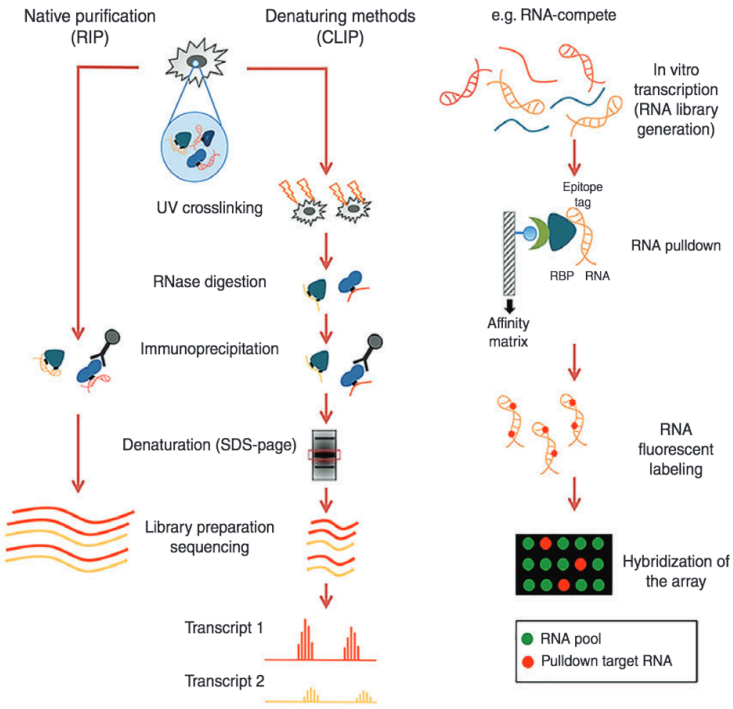


Figure 13. Schematic of protein-centric methods to investigate RNA-protein interactions. *In vivo* approaches include RNA immunoprecipitation (RIP) and a variety of CLIP methods in which a protein of interest is immunoprecipitated and bound RNAs analyzed by sequencing. *In vitro* methods like RNA-compete interrogate an immobilized protein with an RNA library and bound RNAs are detected by microarray hybridization.

In contrast, TRIBE –targets of RNA-binding proteins identified by editing– and RNA tagging are among the methods that do not require protein purification. TRIBE fuses an ADAR family enzyme to the protein of interest that is able to deaminate nearby adenosines (McMahon et al., 2016). Similarly, RNA tagging uses a fused poly(U) polymerase to the RBP that marks bound RNA with poly(U) tails, subsequently identified by 3'-end sequencing (Lapointe, Wilinski, Saunders, & Wickens, 2015).

Existing novel methods in the field will involve engineering of Cas proteins to specifically label RNA or proteins and developing new RNA-protein cross-linking components (Ramanathan et al., 2019). For instance, in TAG-eCLIP, a carboxy-terminal tag is incorporated to the endogenous RBP gene via the

CRISPR-Cas9 system and bypasses the need of specific antibodies for RBP immunoprecipitation (Van Nostrand et al., 2017).

Methods to study global RBP-interactions

RNA interactome capture (RIC) has been used to study the mRNA-binding proteome (mRBPome) in many species including mammalian cell lines, yeast, Drosophila embryos and *Arabidopsis thaliana* (Baltz et al., 2012; Castello et al., 2012; Beckmann et al., 2015b; Sysoev et al., 2016; Marondedze et al., 2016). After irradiation of living cells with UV light to cross-link RNA and proteins that are in direct contact with each other, cells are lysed and poly(A)-RNA purified using oligo(dT))-coupled beads. After extensive washing, the bound RBPs are eluted and coupled to MS measurement. In this setup, RBPs are defined as compared to a control sample that has not been irradiated. RIC data have revealed hundreds of novel RBPs that lack RBDs and are unrelated to RNA biology. Nonetheless, technical variations in RIC protocols –i.e. cross-linking energy, time and RNA labelling– resulted in partially overlapping sets of RNA-binding proteins (Beckmann, 2017)(Figure 14). Thus, experimental variability and technical noise limits the use of RIC, since it has been shown that RIC proteomes can be contaminated with DNA-binding proteins (Conrad et al., 2016). To overcome these limitations, a study from the Hentze laboratory proposed an enhanced RIC protocol (eRIC) that aims to reduce DNA and rRNA contamination that is sensitive enough to enable identification of mRBPome dynamic changes (Perez-Perri et al., 2018).

While RIC reports proteins binding to polyadenylated RNA, RBPs that associate with non-adenylate RNA species cannot be evaluated. Variant approaches that include tRNA, rRNA, small nuclear RNA (snRNA) and small nucleolar RNA (snoRNA) have been proposed (Rinn & Chang, 2012). Urdaneta and co-workers used phenol-toluol (PTex) to isolate RNA based on physicochemical properties and therefore enable the study of the RBPome of non-adenylate RNA as well (Urdaneta et al., 2019). Also, recent development of the TRAPP method based on recovery of RNA-peptide conjugates allows identification of the precise

amino acid at the site of cross-linking (Shchepachev et al., 2019). A recent approach from the Krijgsveld group, successfully applied a method based on acid guanidinium thiocyanate-phenol-chloroform (TRIZOL) purification of cross-linked protein-RNA. TRIZOL is often used to separate RNA from proteins because RNA remains in the aqueous phase and proteins in the organic phase. UV-crosslinked protein-RNA can then be purified from the insoluble interphase, DNase-treated and analyzed by mass spectrometry (Trendel et al., 2019).

RNA-binding proteins repertoire in *Saccharomyces cerevisiae*

In 2010, two independent studies combined protein microarrays and proteomic approaches to uncover RBPs bound to poly(A) RNA and identified about 200 yeast RBPs (Scherrer, Mittal, Janga, & Gerber, 2010; Tsvetanova et al., 2010). Moreover, different RIC studies revealed partially overlapping sets of RBPs due to technical differences among the protocols (Figure 13) (Beckmann et al., 2019; Matia-gonzález et al., 2015; S. F. Mitchell et al., 2012). Similar to other species, the resulting mRBPome also includes proteins with no annotated RBDs, corresponding to known enzymes, already suggesting the existence of a larger amount of RBPs as previously anticipated.

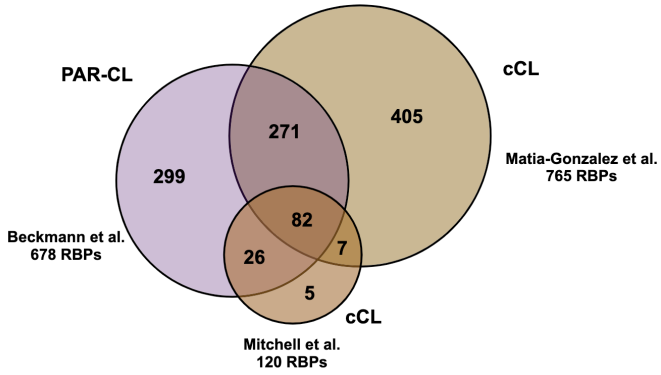


Figure 13. RBP repertoire in yeast. Venn diagram showing RBPs identified by different studies in yeast using either photoactivatable-ribonucleoside-enhanced crosslinking (PAR-CL) or 254 nm UV light crosslinking (cCL).

Indeed RNA-binding proteins play essential roles in many cellular processes and the RBP repertoire in several organisms has been addressed from diferent

experimental perspectives. Here, we investigate the set of RBPs binding to individual RNA structures that show evolutionary conservation.

II. AIMS OF THE THESIS

AIMS OF THE THESIS

In my PhD I applied quantitative proteomics to investigate organismal development and in different interactomics studies to explore proteins bound to peptide modifications, telomeric DNA and structured RNA molecules.

In Chapter III - The Developmental Proteome of Drosophila melanogaster - we aimed at:

- Cataloguing a comprehensive dataset of proteins expressed throughout the complete life cycle of the fruit fly.

- Provide a proteome dataset with high temporal resolution during the tightly regulated process of embryogenesis.

- Use our generated datasets to explore global RNA-protein correlations, functionally characterize some developmentally regulated proteins as well as identify maternally loaded, gender- and age-specific proteins.

In Chapter IV - The RNA Fold Interactome of Evolutionary Conserved RNA Structures in *S. cerevisiae* - we worked towards:

- Applying a systematic quantitative proteomics screen to unravel proteins binding to a set of evolutionary conserved RNA structures.

- Employing orthogonal techniques to demonstrate *in vivo* binding of selected protein-RNA pairs

- Exploring functional consequences of some interactions using a reporter screen and SILAC experiments and using genetic data integration to gain biological insights on the protein-RNA pairs.

III. The developmental proteome of Drosophila melanogaster

Genome Res. 2017 Jul;27(7):1273-1285.doi:10.1101/gr.213694.116.

The developmental proteome of *Drosophila melanogaster*

Nuria Casas-Vila[1,8], Alina Bluhm[1,8], Sergi Sayols[2,8], Nadja Dinges[3], Mario Dejung[4], Tina Altenhein[5], Dennis Kappei[6], Benjamin Altenhein[5,7], Jean-Yves Roignant[3] and Falk Butter[1,9]

[1] Quantitative Proteomics, Institute of Molecular Biology (IMB), Mainz, Germany
[2] Bioinformatics Core Facility, Institute of Molecular Biology (IMB), Mainz, Germany
[3] RNA Epigenetics, Institute of Molecular Biology (IMB), Mainz, Germany
[4] Proteomics Core Facility, Institute of Molecular Biology (IMB), Mainz, Germany
[5] Institute of Genetics, Johannes Gutenberg University (JGU), Mainz, Germany
[6] Cancer Science Institute of Singapore, National University of Singapore (NUS), Singapore
[7] Institute of Zoology, University of Cologne, Cologne, Germany
[8] Equal contribution
[9] To whom correspondence should be addressed: f.butter@imb.de; Institute of Molecular Biology (IMB), Ackermannweg 4, 55128 Mainz

Keywords: Drosophila melanogaster, gene regulation, development, embryogenesis, proteomics, systems biology

ABSTRACT

Drosophila melanogaster is a widely used genetic model organism in developmental biology. While this model organism has been intensively studied at RNA level, a comprehensive proteomic study covering the complete life cycle is still missing. Here, we apply label-free quantitative proteomics to explore proteome remodeling across Drosophila's life cycle, resulting in 7,952 proteins, and provide a high temporal-resolved embryogenesis proteome of 5,458 proteins. Our proteome data enabled us to monitor isoform-specific expression of 34 genes during development, to identify the pseudogene Cyp9fpsi as a protein-coding gene and to obtain evidence of 268 small proteins. Moreover, the comparison with available transcriptomic data uncovered examples of poor correlation between mRNA and protein, underscoring the importance of proteomics to study developmental progression. Data integration of our embryogenesis proteome with tissue-specific data revealed spatial and temporal information for further functional studies of yet uncharacterized proteins. Overall, our high-resolution proteomes provide a powerful resource and can be explored in detail in our interactive web interface.

INTRODUCTION

Drosophila melanogaster is among the best-described model organisms for development and ageing. During its life cycle, it progresses through well-defined stages including embryo, larva, pupa and adult, undergoing a complete phenotypic metamorphosis (Lawrence 1992). These transitions are based on tightly regulated gene expression at the transcriptional, epigenetic and translational level. Currently, most developmental gene expression studies in Drosophila rely on in situ hybridization of RNA (Lécuyer et al. 2007; Tomancak et al. 2007), transcriptome analysis using large-scale microarray/RNA-seq data sets (Brown et al. 2014; Chintapalli et al. 2007; Graveley et al. 2011; Kalinka et al. 2010) or a combination of both (Jambor et al. 2015). However, mRNAs are further translated into proteins, which perform the actual cellular functions. It has been shown in multiple species such as Saccharomyces cerevisiae (Griffin

et al. 2002), Trypanosoma brucei (Butter et al. 2013), Caenorhabditis elegans (Grün et al. 2014), human (Schwanhäusser et al. 2011) as well as in Drosophila melanogaster (Bonaldi et al. 2008) that transcript levels are only a moderate predictor for protein expression as they do not account for posttranscriptional processes such as translational regulation or protein stability (Liu et al. 2016; Vogel and Marcotte 2012). Recently, this has also been addressed with a developmental perspective in Caenorhabditis elegans (Grün et al. 2014), Xenopus laevis (Peshkin et al. 2015) and Trypanosoma brucei (Dejung et al. 2016), but not yet in Drosophila.

The number of fly proteins with available antibodies increased in the last decade from around 450 (Adams et al. 2000) to 1,586 (listed in FlyBase version 6.01), but still covers only a small fraction of expressed genes. To accelerate protein studies in Drosophila, several tagging strategies were devised. Around 100 genes have been fused with a GFP using piggyBac transposition (Morin et al. 2001) and 400 GFP-tagged fly lines have been established using MiMICs (Minos Mediated Integration Cassette) to permit systematic protein investigations (Nagarkar-Jaiswal et al. 2015). In an alternative approach, BAC TransgeneOmics allowed the creation of 880 lines and the systematic study of 207 GFP-tagged fly proteins (Sarov et al. 2016). In principle, all protein-coding genes can be investigated, but this requires the establishment of a line for each protein. Additionally, a putative caveat of tagging strategies is altered protein behavior like mislocalization, changes in protein stability or a dominant negative regulatory effect (Margolin 2012).

The fruit fly is one of the model species investigated by modENCODE (The modENCODE consortium et al. 2010) and thus several large data sets are available for mapping histone modifications (Kharchenko et al. 2011), global RNA levels during development (Graveley et al. 2011) and tissue-specific splicing (Brown et al. 2014). In contrast, proteomic studies in Drosophila have been restricted to certain developmental stages. For example, changes in the proteome during ageing from eclosure to 60 days old flies (Sowell et al. 2007), the adult itself (Sury et al. 2010; Xing et al. 2014), larva and pupa (Chang et al.

2013), the embryo (Fabre et al. 2016) and the oocyte-to-embryo transition (Kronja et al. 2014) have been investigated. However, these studies have relatively low proteome coverage (around 2,000 proteins), do not cover the complete developmental process and are not directly comparable because of technical differences.

Applying label-free quantitative proteomics (Cox et al. 2014), we here measured protein expression throughout the Drosophila life cycle with a coverage of 7,952 proteins to provide insight into proteome remodeling. With embryogenesis being a focus in Drosophila developmental studies, we amended the life cycle proteome with an embryogenesis proteome of 5,458 proteins with high temporal resolution. Finally, data integration with tissue-specific (Lécuyer et al. 2007) and developmental transcriptomic studies (Graveley et al. 2011) allows investigating the importance of spatial and translational regulation.

RESULTS

Proteomics screen of the life cycle

We collected whole animal samples at 15 representative time points during the Drosophila life cycle (Fig. 1A). The embryonic time points were chosen according to major stages of embryonic development: prior to zygotic gene activation (0-2h, E02), gastrulation (4-6h, E06), organogenesis (10-12h, E10) and the late stages of embryogenesis (18-20h, E20). For larva, the three different instar larva (L1, L2 and early L3) and a late stage (L3 crawling larva) were examined. Pupae were collected daily starting with the white pupa and, for adults, the virgin males and females (up to 4h after eclosure) as well as one week old animals of each sex were chosen. All samples were collected as biological quadruplicates and processed by mechanical disruption with a universal protein extraction protocol. For each replicate, a five hour mass spectrometry (MS) run was used, resulting in 340 hours of measurement (68 MS runs). We searched the resulting 8 million MS/MS spectra against a Saccharomyces cerevisiae and Drosophila melanogaster database using the MaxQuant software suite (Cox and Mann 2008). Overall, we identified 9,627

protein groups (a protein group contains proteins indistinguishable by the peptides that were identified) with 144,067 unique peptide sequences at a FDR<0.01. This number includes 1,078 yeast and 8,549 Drosophila protein groups (Supplemental Fig. S1A). The identification of yeast proteins is nearly exclusively restricted to the larval stages where it is a food source (Supplemental Fig. S1B). The number of 8,549 identified fly proteins is comparable to a previous in-depth measurement of multiple sources of Drosophila material reaching 9,124 proteins (Brunner et al. 2007). After filtering for robust detection in at least two replicates of any time point, we performed our subsequent analysis on a set of 7,952 protein groups (Fig. 1B and Supplemental Table S1). Developmental processes are tightly regulated and thus highly reproducible in each organism. Nevertheless, to visualize biological variability of this process, we performed correlation and principal component analysis (PCA).

To increase quantitation reliability, all label-free quantitation (LFQ) values were solely based on unique peptide intensities for each protein group. Despite that our replicates are originating from different egg-laying events, being processed independently and measured several days apart on the mass spectrometer, we find a very high correlation within the time points (R = 0.84-0.98) (Supplemental Fig. S1C) and clear formation of clusters in PCA (Fig. 1C). These findings demonstrate a very high reproducibility of our experimental conditions from the biological system to the mass spectrometry measurement.

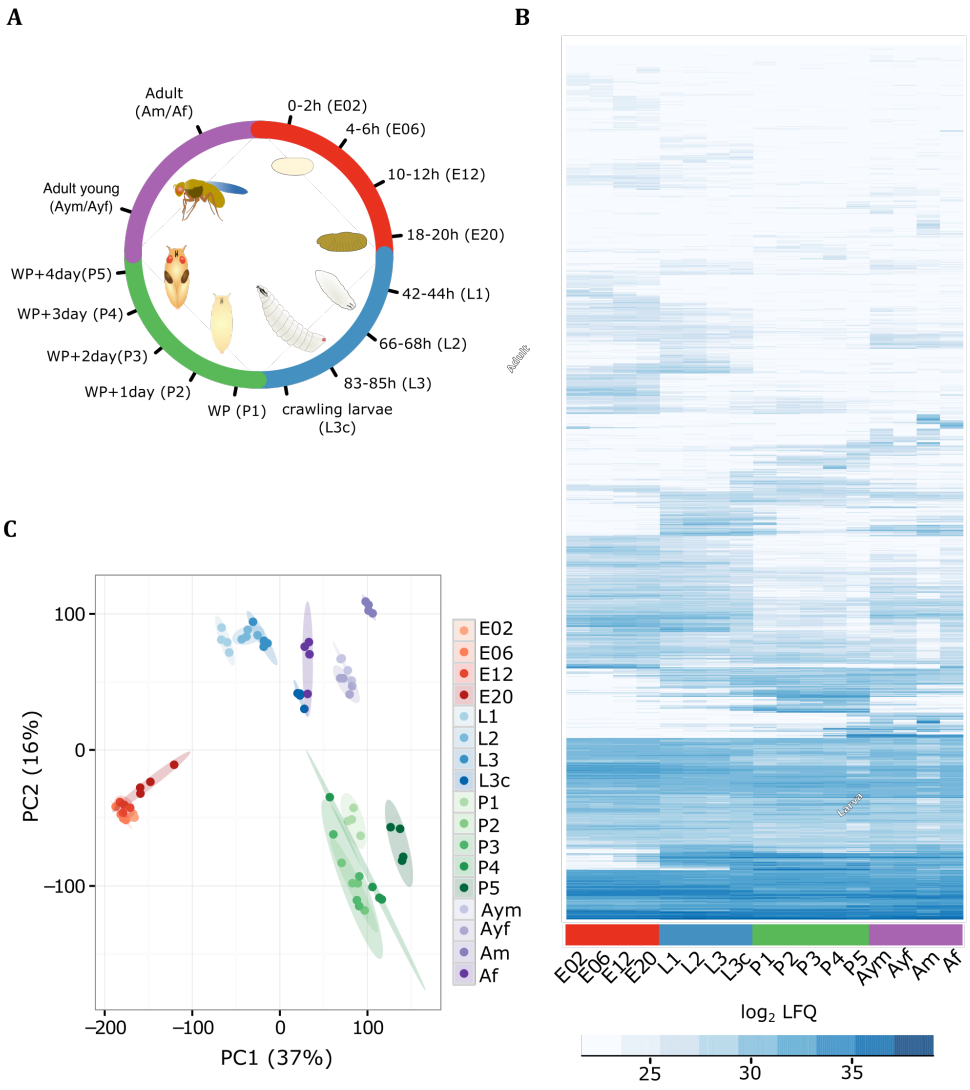


Fig. 1. Drosophila developmental life cycle proteome.
(A) Scheme depicting the collected time points throughout the four major metamorphic stages of Drosophila (embryo [red], larva [blue], pupa [green] and adult [violet]). WP: white pupa, L3c: crawling third instar larva. (B) Heat map of log2 LFQ values of the 7,952 protein groups quantified during fly development. (C) Visualization of the first two principal components separating samples according to their developmental stage. The biological replicates are indicated in the same color with elliptic areas representing the standard error of the two depicted components.

Core proteome and protein expression dynamics

To identify a core proteome, i.e. proteins detected at all stages of development, we grouped the proteins according to their presence in the four major stages of the life cycle (Fig. 2A). We found 4,627 proteins groups, more than half of our proteome, to be detectable in all stages. To obtain an overview of the functionality of these continuously expressed proteins, we performed gene ontology (GO) annotation enrichment analysis and reduced the GO term complexity to uncover major descriptors (Fig. 2B). As expected, our core proteome is enriched for metabolic and cellular processes describing the basic activities of any cellular system, exemplified by covering all known proteins for such essential processes as tRNA aminoacetylation, endosome transport via multivesicular body sorting pathway, cell junction maintenance, nuclear pore organization and ribosome assembly (Fig. 2B and Supplemental Fig. S2A, Supplemental Table S2). We also analyzed developmental expression dynamics for all proteins with an averaged abundance above the detection limit, log2 LFQ intensity >25 (Fig. 2C, Supplemental Table S3). We additionally applied a Gini coefficient filter of 0.1, which divided our proteome into 1,386 stably expressed proteins throughout life cycle and 1,978 differentially expressed proteins. Consistent with a previous developmental study in Xenopus, we see that the dynamicity decreases with protein abundance (Peshkin et al. 2015) (Fig. 2C). We show examples of highly dynamic and stably expressed proteins (Fig. 2D, Supplemental Fig. S2D). The stable proteins include the widely accepted loading controls: tubulins, actins, heat-shock proteins, Gapdh1, Gapdh2 and Vinculin.

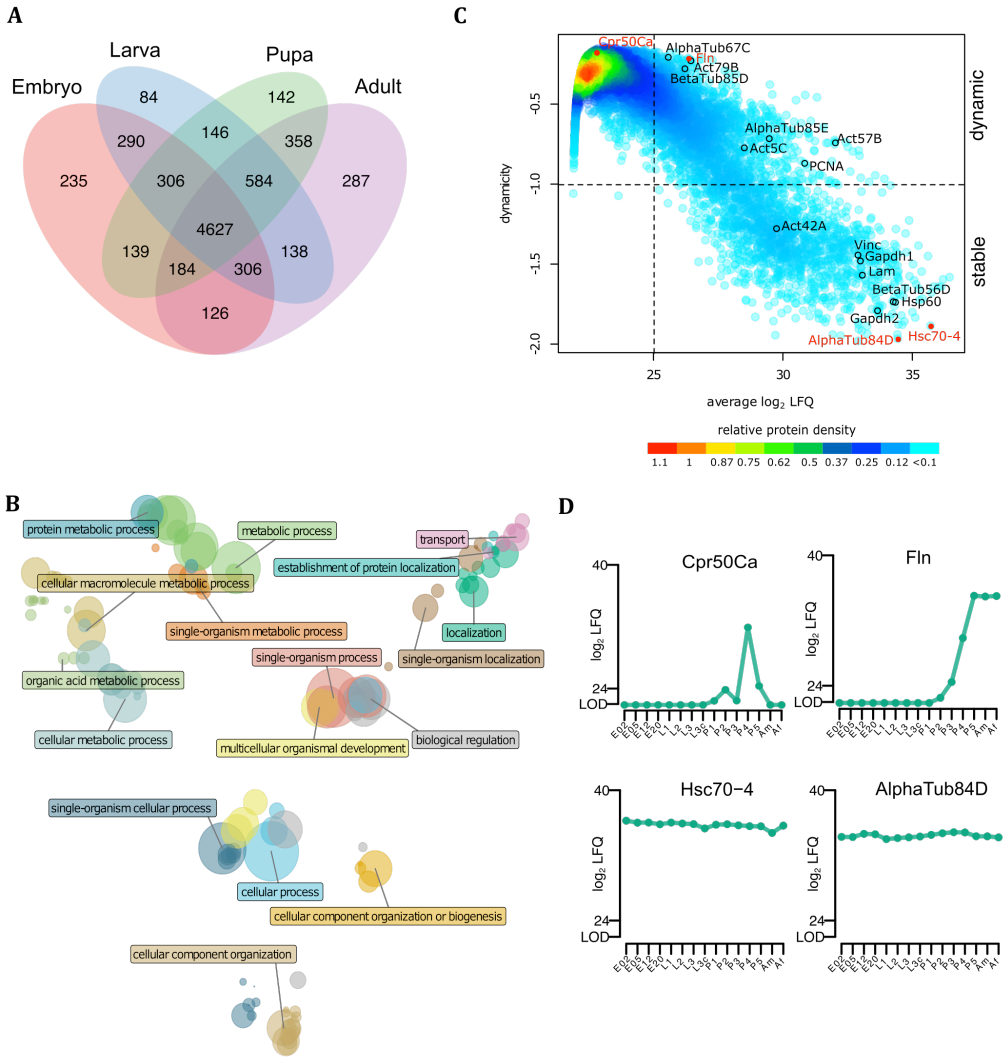


Fig. 2. Characteristics of the developmental proteome.
(A) Overlap of quantified protein groups between developmental stages results in a core proteome of 4,627 proteins. (B) Clusters of enriched GO terms obtained from the core proteome are plotted in a coordinate system defined by the first two dimensions of a multidimensional scaling according to their similarity scores. The color of the circle represents the GO cluster with a representative term highlighted. The diameter of the circle is proportional to the size of the GO category. (C) The density plot relates protein abundance with a dynamicity score during developmental protein expression (log10 transformed Gini index). In the lower-right quadrant, highly stable proteins are represented, while the upper-right quadrant contains proteins with changing expression levels during development. (D) Expression profiles for two highly dynamic (upper panel) and two stably expressed (lower panel) proteins highlighted in red in the dynamicity plot.

Developmental expression profiles of highly abundant proteins

We first characterized the 100 most abundant proteins per stage, comprising around 10% of the total protein mass (Supplemental Table S1). Among proteins with the highest LFQ values, we find ribosomal proteins, being especially prevalent in the top 100 list during embryogenesis, a phase of rapid cell proliferation. The fly uses different storage proteins at specific developmental stages: yolk proteins (Yp1, Yp2 and Yp3) in embryogenesis and Lsp proteins whose protein abundance rises drastically in L3. Among these highly abundant proteins there are several preliminary annotated genes that are not further characterized. CG1850, representing the most abundant protein in the pupal stage, shares a small stretch of similarity to the cuticular protein Cpr72Eb (BLAST E-value: 0.019, Supplemental Fig. S2B). Interestingly, some other highly expressed computed genes (CG) also show similar protein expression patterns to well-studied cuticular proteins like Cpr72Ea (CG1850 and CG13023), Cpr64Aa and Cpr64Ac (CG34461 and CG42323) and Cpr66D (CG16886 and CG30101). While thus far we looked at the most highly expressed 100 proteins, our proteome can be interrogated to reveal the temporal expression pattern of any quantified protein.

Proteome remodelling throughout the life cycle

Our proteome covers a dynamic range of more than 6 orders of magnitude showing expression changes of individual proteins of more than 100,000 fold (Supplemental Fig. S2C). We interrogated our data set for stage-specific proteins by applying ANOVA (FDR<0.01) on the log2 LFQ values (Fig. 3A). The majority of these 1,535 differentially regulated protein groups are found in adult flies (556), followed by embryos (473), pupae (317) and larvae (189). To connect the proteome differences to stage-specific biological functions, we performed GO enrichment analysis on clustered protein expression profiles (Fig. 3A, Supplemental Tables S4 and S5). The most enriched GO terms during embryogenesis include mitotic cell cycle regulation and nuclear division represented by cyclins (CycE, CycA, CycB) and developmental kinases, such as

Loki (Lok), Greatwall (Gwl) and Grapes (Grp). By this clustering, we were able to separate an early and late embryogenesis phase (Fig. 3B). The early phase (0-6h) is characterized by high expression of proteins involved in cytoskeleton organization (Dgt4, AlphaTub67C and GammaTub37C), microtubule binding proteins (Mars and Wee Augmin (Wac)) as well as the classical examples Bicaudal C (BicC) and Cup, important in translational regulation of the oskar mRNA. In contrast, proteins involved in tissue morphogenesis such as Bazooka (Baz), Fat (Ft), Ribbon (Rib) and Tramtrack (Ttk) are upregulated in later phases (12-20h). Stage-specific proteins in larvae and pupae include expected structural constituents of the chitin-based cuticle: Lcp, Tweedle (Twd) and cuticular proteins. Intriguingly, several proteins that are highly upregulated only at a single pupal stage, like CG13376, CG13082 and CG42449, are poorly characterized (Fig. 3B and Supplemental Fig. S3A). In the adult, odorant-binding proteins (Obp83b and Obp57a), proteins involved in light perception and phototransduction (Arr1 and Arr2) and the retinal degeneration protein A (RdgA) show strong expression, consistent with the adult fly having a fully developed light sensory system. Also proteins involved in muscle contraction like flightin (Fln) and Eaat1, increase their expression 100-fold in adult stages (Fig. 3B).

Overall, our data is in agreement with previously published studies and connects protein expression with well-described morphological changes during Drosophila development. Therefore, our screen defines the developmental stage to study molecular or phenotypic effects of yet uncharacterized proteins. All protein profiles can be interrogated using the interactive web interface (http://www.butterlab.org/flydev).

Developmentally regulated functions: ecdysone-induced proteins and cuticle formation

The regulation of molting by endogenous 20-hydroxyecdysone (20E) is a prototype example of hormonal gene regulation pathways in insects (Yamanaka et al. 2013). Previous microarray studies focused on 20E-induced gene regulation of mRNA transcripts between L3 larval stage and 12h after puparium formation (Beckstead et al. 2005; Gonsalves et al. 2011). However, for the ecdysone-induced gene family 71E (Eig71E), we find intriguing differences between the expression profiles of mRNA and protein in pupae. Messenger RNA expression is detectable in three different waves: Eig71Ee spikes at L3c, another group represented by Eig71Ed at P1 and a later group represented by Eig71Ek at P2 (Graveley et al. 2011) (Fig. 3D and Supplemental Fig. S3B).

While the mRNA is detectable only in early pupal stages, the corresponding Eig71E proteins show prolonged high expression levels until P5 (Fig. 3C and 3D). Likewise, second puff genes display a similar transcriptome versus proteome pattern. A 1000-fold upregulation of glue proteins (Sgs5, Sgs7 and Sgs8) at late L3 concordant to the detection of their mRNA in a narrow window of circa 24h between crawling L3 and P1 (Beckstead et al. 2005), is followed by the presence of the protein in all pupal stages (Fig. 3D and Supplemental Fig. S3C). Our data shows that for selected puff proteins, protein stability is the major determinant of their expression patterns during development. In contrast, in a high number of cases, we detect the protein at a single time point while the RNA is detectable at multiple time points (Fig. 3E). In the aforementioned cases, protein levels cannot be directly predicted by transcriptomics, which demonstrates the necessity of proteome data for studying fly development.

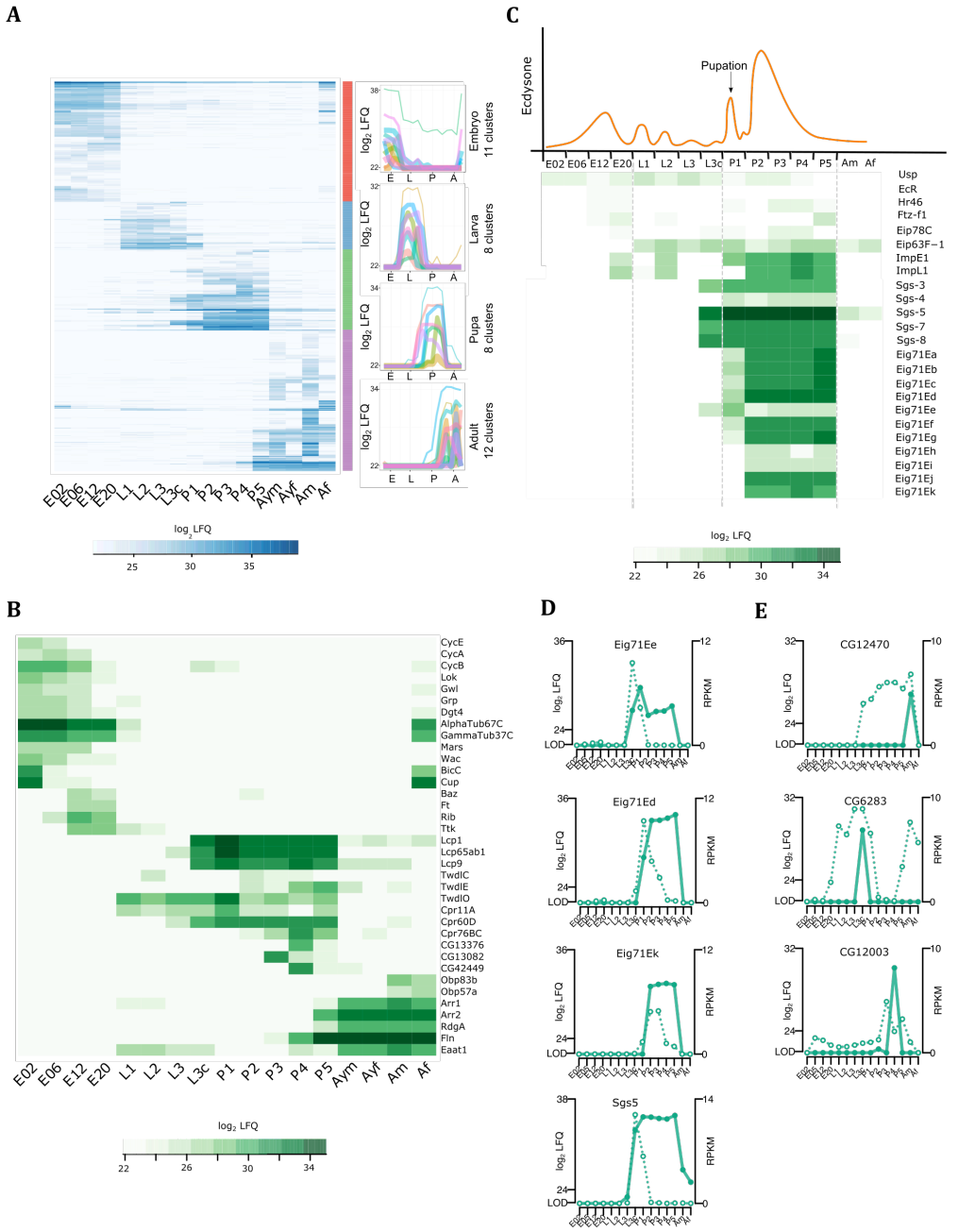


Fig. 3. Stage-specific proteins and ecdysone-induced developmental regulation.
(A) Heat map showing 1,535 protein groups found to be differentially (ANOVA, FDR<0.01) regulated during the life cycle. These protein groups were clustered into up to 12 stage-specific profiles. Average profiles of the individual clusters for each developmental stage are shown. B) Heat map showing log2 LFQ abundance of proteins with stage-specific expression

profiles discussed in the text. (C) Schematic representation of ecdysone pulses during fly development (upper panel) and heat map of log2 LFQ expression levels of selected proteins of 20-hydroxyecdysone regulated genes (lower panel). (D) For the Eig71E and Sgs gene family, RNA expression profiles (dotted line) differ from protein levels (solid line) during the pupal phase demonstrating prolonged protein stability. (E) Three examples showing single protein expression burst, but more broadly detectable RNA indicating more tightly controlled protein expression.

Comparison of gender-specific protein patterns in adult flies

Gender-specific proteins are of high interest and have already been investigated by several proteomics studies (Dorus et al. 2006; Sury et al. 2010; Takemori and Yamamoto 2009; Wasbrough et al. 2010). To benchmark our label-free quantitative approach we compared our adult time point to the published SILAC data set (Sury et al. 2010) and found a high overlap of gender-specific proteins (R=0.84, Supplemental Fig. S4A), showing that our developmental proteome recapitulates previous studies that are more specialized. To identify gender-specific proteins, we defined a 4-fold expression difference with a p-value below 0.01 between male and female flies (one week old) and found 308 male- and 374 female-specific proteins (Fig. 4A, Supplemental Table S6). The 308 male proteins include Tektin-A and Tektin-C as sperm-specific flagellar proteins (Amos 2008), several less characterized genes known to be expressed in fly testes and seminal vesicles (Dorus et al. 2006; Takemori and Yamamoto 2009) and some proteins functioning in male development like Lectin-46Ca, Lectin46Cb and Lectin-30A. For some proteins like Hsp60B, Hsp60C, the male fertility factor Kl-5 as well as Aquarius (Aqus) and Antares (Antr), an essential role in sperm development or sperm storage has already been demonstrated. The list of 374 female-specific proteins include vitelline membrane (Vm32E) and chorion proteins (Cp15, Cp18, Cp36), which are important for eggshell assembly, the vitellogenins (Yp1, Yp2 and Yp3) and the fatty acid desaturase Fad2. Additionally, our developmental proteome allows the investigation of young flies, which were collected as virgins within 4h after eclosure (Supplemental Table S7). The majority of proteins are equally expressed in both genders (Supplemental Fig. S4B). While we detect only 21

female-specific proteins in young flies, there are 155 proteins with higher expression in its male counterpart (Fig. 4B, Supplemental Fig. S4C and S4D).

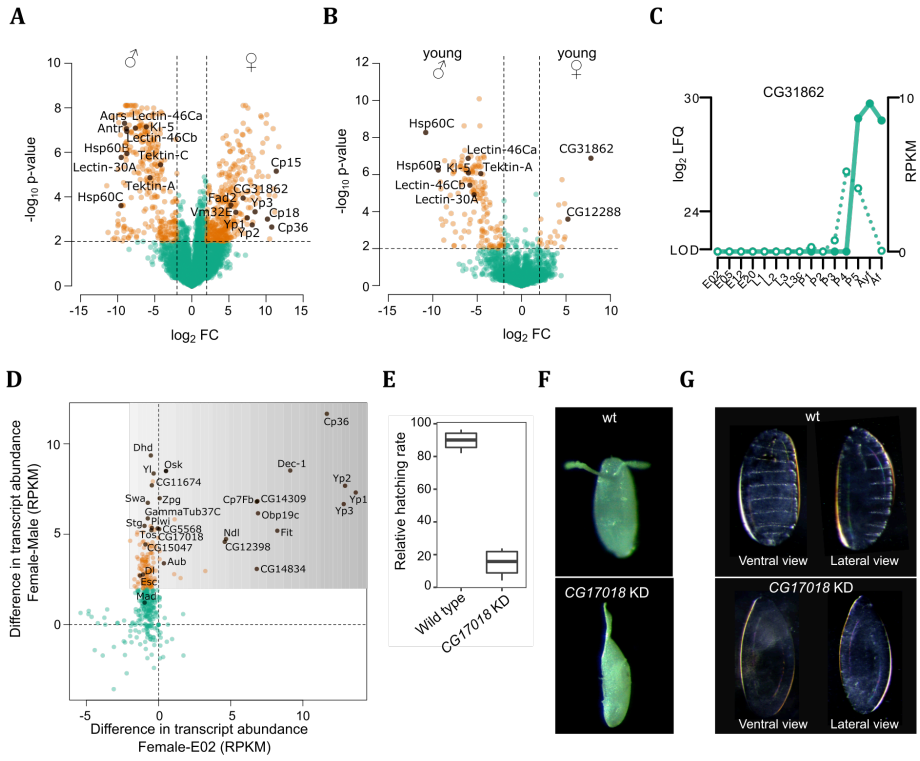


Fig. 4. Gender-specific proteome and maternally loaded proteins.
(A) Volcano plot comparing protein expression levels between one-week-old male and female flies. Candidates discussed in the text are highlighted (filled black circle). Dashed lines indicate a 4-fold expression difference with p<0.01. (B) Volcano plot comparing protein expression levels between young male and female flies (less than 4 hours old after eclosure) shows very few female-specific proteins. Candidates discussed in the text are highlighted (filled black circle). Dashed lines indicate a 4-fold expression difference with p<0.01. (C) Developmental expression profile of the female-specific protein CG31862 shows detection of mRNA (dotted line) in late pupal stage, while the protein (solid line) is also found in female flies. (D) Integration of mRNA levels with embryo-specific proteins allows identifying maternally loaded proteins. The mRNA levels of the adult female flies compared to embryos (x-axis) and males (y-axis) distinguishes cases in which either both, the mRNA and protein (x=0, y>2) or only the protein (darker shaded area) is maternally loaded. (E) Relative embryonic hatching rate (four biological replicates) of CG17018 knockdown embryos compared to wild type. (F) Image of representative wild type and the CG17018 knockdown embryo with fused dorsal appendages. (G) Cuticle preparation of embryos revealed absence of denticle belts patterning in the CG17018 knockdown line.

In agreement with this observation, a previous transcriptomic study showed upregulation of genes in female flies after mating, suggestively triggered by

sperm and seminal fluid proteins (McGraw et al. 2008). In contrast, the majority of male-specific proteins is already present in young male flies prior to mating (Fig. 4B, Supplemental Fig. S4D). Interestingly, the only two proteins with more than 30-fold upregulation in virgin females compared males are not characterized: CG31862 and CG12288. Noteworthy, CG31862 is found in P5 and shows continuously high protein level, while its RNA expression is restricted to the late pupal phase (Fig. 4C).

Maternally loaded proteins

While there is ample knowledge about maternally loaded RNA in Drosophila embryos (Tadros and Lipshitz 2005), no systematic analysis for maternally loaded proteins has been conducted yet. We interrogated our data for proteins enriched during embryogenesis whose RNA levels were higher in adult females compared to adult males. Among this subset of likely maternally loaded material should be candidates that have a functional importance during early development. In most cases, protein and mRNA are present in 2h old embryos, suggesting that both are maternally loaded (Fig. 4D and Supplemental Table S8). These include well-known examples such as Oskar (Osk), String (Stg), Piwi, Aubergine (Aub), Extra sexcombs (Esc), Dorsal (Dl), Mothers against dpp (Mad) and Swallow (Swa) (Chao et al. 1991; Edgar and Datar 1996; Luschnig et al. 2004; Mani et al. 2014; Simmons et al. 2010). However, also yet undescribed candidates like CG11674, CG5568, CG17018, CG15047, Zpg, GammaTub37C and Tosca found in this set represent interesting candidates with a putative role in oogenesis and early embryogenesis. In order to investigate potential germline-specific functions of these candidates we performed RNAi mediated-knockdown using the driver nanos-GAL4 and specific transgenic lines expressing double–stranded RNA from inverted repeats (shRNAs). Germline-specific expression of two independent shRNAs targeting CG17018 RNA revealed drastic effects on the embryonic hatching rate. While the number of laid eggs was unaffected (Supplemental Fig. S4E), hatching was reduced by almost 80% (Fig. 4E). Besides, approximately 30% of unhatched eggs displayed

defective dorsal appendages that are fused (Fig. 4F). Cuticle preparations showed that CG17018 knockdown embryos miss the denticle belts, revealing an absence of patterning at early stages (Fig. 4G). To note, CG17018 knockdown ovaries were indistinguishable from wild type ones, as we could not detect any obvious morphological or differentiation defects (Supplemental Fig. S4F). Taken together, our findings imply a critical role of CG17018 during early embryogenesis.

Furthermore, our proteomic data set allows a comprehensive classification of maternally loaded proteins when the RNA is not present. The most prominent proteins include the major egg yolk vitellogenins (Yp1, Yp2 and Yp3), Dec-1, Cp36 and Cp7Fb as part of the chorion, the oxidoreductase family member CG12398 for which a role in vitelline membrane formation has been previously suggested (Fakhouri et al. 2006), the serine protease Nudel (Ndl), the sensor protein Obp19c, the female-specific protein Fit as well as two uncharacterized candidates, CG14309 and CG14834 (Fig. 4D and Supplemental Table S8).

Small proteins in the developmental proteome

Recently, there has been an increased interest in small proteins and translated small ORF (smORF) with up to 100 amino acids (Ramamurthi and Storz 2014) as their protein coding potential is difficult to assess bioinformatically (Ladoukakis et al. 2011). These small proteins localize to specific subcellular compartments and perform cellular functions as any other protein (Magny et al. 2013). Our data set detects 268 small proteins (Fig. 5A) of which 84% have two or more unique peptides and temporal expression information (Supplemental Fig. S5A and Supplemental Table S1). This number is similar to a previous investigation using ribosome profiling (Aspden et al. 2014), demonstrating that mass spectrometry-based proteomics is on par with next generation sequencing approaches to detect translation of small proteins.

Peptides originating from non-coding regions of the genome

Peptides originating from putative non-coding regions have been reported in diverse organisms. Therefore, we re-analyzed our data including ncRNA sequences from FlyBase, which we *in silico* translated for open reading frames of at least 20 amino acids. Overall, we identified 29 putative proteins that unambiguously map to non-translated transcripts at a FDR<0.01 (Supplemental Table S9). Due to short open reading frames of these small proteins, we usually detect a single peptide per transcript. However, only two of these ncRNA-derived peptides showed a good MS2 fragmentation pattern and were independently identified with more than 10 different MS/MS spectra in several replicates and time points. One of these, FBtr0340701 has also been found in a control experiment using human cell lysate (data not shown), classifying it as a false positive identification originating from a contaminant. The only remaining peptide with strong evidence of identification matches to CR43476 (Fig. 5B).

Other genes classified as non-expressed are pseudogenes. These genes have mutations in their promoter regions or other functional elements that make their expression unlikely (Harrison et al. 2003). We checked for protein evidences of the 2,902 reported pseudogenes (FlyBase 6.01), and found 9 protein groups in our dataset to include peptides unambiguously mapping to pseudogenes. Whereas most of these proteins are represented by a single peptide (Supplemental Table S10), the most prominent hit, FBtr0082602 encoding Cyp9f3psi, is supported by 23 peptides including 5 unique sequences. The measured peptides match to the N-terminal and C-terminal regions, demonstrating that the complete pseudogene is most likely translated (Fig. 5C). Furthermore, Cyp9f3psi and Cyp9f2 present distinct expression patterns, further indicating that despite their close genomic vicinity they are differently regulated during development (Supplemental Fig. S5B).

Despite an extremely low expression of peptides originating from ncRNA transcripts, only very few detected peptides map to non-coding regions of the genome, illustrating a low false discovery rate in our screen and a carefully

curated gene annotation of the Drosophila melanogaster genome (Matthews et al. 2015).

Highly temporal-resolved embryogenesis proteome

Being intensely studied, we were particularly interested in proteome changes during embryogenesis. To investigate the process in a high time-resolved and systematic manner, we collected whole embryos at narrow intervals: every single hour after egg laying for up to 6h and then every two hours until 20h (Fig. 6A). These 14 time points were also measured in four independent biological replicates to account for technical, biological and environmental variation. To control for our collections, we staged embryos of selected time points by morphology and Engrailed antibody staining (Campos-Ortega and Hartenstein 1997) (Supplemental Fig. S6A).

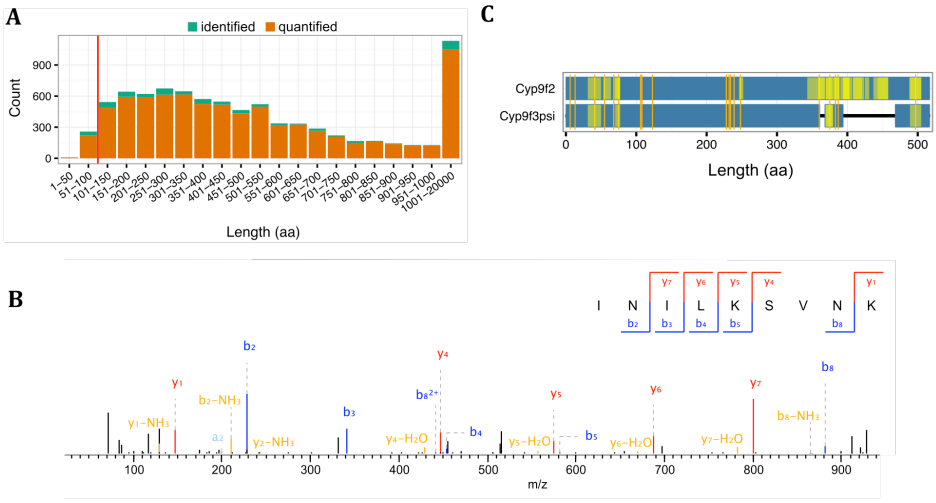


Fig. 5. Small proteins and peptides from non-coding regions of the genome.
(A) Protein length distribution of identified (green, not enough quantitation values) and quantified (orange) protein groups of the life cycle proteome. Most proteins have quantitation values (>90 percent) and this fraction only marginally depends on protein length. The red line demarcates the fraction of 268 small proteins (<100 aa). (B) Representative MS/MS spectrum with annotated b- and y-ions of the peptide INILKSVNK(2+) from the putative non-coding gene CR43476. (C) Sequence comparison of Cyp9f2 and the "pseudogene" product Cyp9f3psi with amino acid substitution between both proteins marked in orange. Coverage of peptides for either protein is shown (yellow, more intense regions have overlapping peptides).

Protein expression levels were determined using label-free quantitation based on unique peptides provided by MaxLFQ (Cox et al. 2014). We detected 6,487 expressed protein groups of which 5,458 were quantified in at least two replicates of any time point (Supplemental Table S11). PCA revealed that embryo stages correlate well with our collected time points (R=0.93), showing a developmental progression through embryogenesis (Fig. 6A, Supplemental Fig. S6C). Noteworthy, all four independent biological replicates show very high reproducibility (R=0.92-0.96) (Supplemental Fig. S6B and S6C). We also validated the expression profiles of seven protein by immunostaining with antibodies against endogenous proteins (Fig. 6C and Supplemental Fig. S6D).

Expression profiles during embryogenesis

We analyzed the time course data using a multivariate empirical Bayes approach and identified 1,644 protein groups with differential expression during embryogenesis (Fig. 6B). To obtain a functional overview on the embryogenesis process, we performed GO enrichment analysis on this set of differentially expressed proteins. Based on this analysis, we observed enrichment of terms related to very early embryogenesis cellular processes (0-1h), such as zygotic determination of anterior/posterior axis and syncytial blastoderm mitotic cell cycle (Fig. 6D). Additionally, proteins involved in ribosome biogenesis upregulate at 2h to initiate active translation concomitant with zygotic gene activation (ZGA) starting at 2h. We also noted high enrichment of proteins involved in cell cycle and cytoskeleton organization during early phases of embryogenesis (2-3h). While proteins involved in nervous system development are highly present at 3h, muscle structure development proteins are more prominent later in embryogenesis at 14h.

As an alternative approach to analyze the data, we automatically clustered the differentially expressed proteins with similar temporal profiles, resulting in 70 distinct clusters and performed GO enrichment analysis on these clusters (Fig. 6E, Supplemental Table S12 and S13). As a result, the known embryonic developmental program can be followed by temporal alignment of individual

clusters (Fig. 6E), possibly hinting at putative functions of not yet characterized proteins.

Integrating the developmental proteome and spatial expression

To integrate spatial information, we fused our proteome profiles with tissue-specific RNA expression data from fluorescence in situ hybridizations (Lecuyer et al., 2007). We chose muscle development to highlight the value provided by the merged data. In the muscle-specific clusters (Fig. 6F, Supplemental Table S14 and S15), we noted upregulation of proteins involved in muscle development such as Mlc2, Mp20 and Mlp60A (Sandmann et al. 2006) at 14h. Later in embryogenesis (20h), we found high expression of Eaat1 and EcR, which control muscle contraction at larval stages. Furthermore, this data integration allowed us to identify similarly expressed, not yet characterized proteins (CG1674, CG6040 and CG15022), shown to localize in muscle tissue, suggesting a role in muscle development. In order to test this hypothesis, we performed RNAi mediated-knockdown of two candidates. Remarkably, mesodermal knockdown of either CG1674 or CG6040 severely affects locomotion behavior of adult flies (Fig. 6G and Supplemental Fig. S6E). Likewise, the complete CG6040 loss-of-function
produces viable flies that display similar climbing defects, confirming the specificity of the RNAi phenotype. Importantly, neuronal knockdown of both genes did not impair locomotion performance, supporting their muscle-specific functions.

We next performed in situ hybridization on embryos and observed a strong enrichment of CG1674 mRNA in muscle tissue, more specifically in somatic and pharyngeal muscles, whereas CG6040 exhibits moderate ubiquitous expression (Supplemental Fig. S6F). Altogether, our findings strongly suggest a muscular function for CG1674 and CG6040. However, further investigations will be required to investigate their specific role in muscle development.

Alternatively, other tissue data can be inspected for biological insights. The analysis of the central nervous system (CNS) revealed an upregulation of

proteins involved in neural development (Roughest (Rst), Smooth (Sm) and Erect wing (Ewg)) at 8-12h and in synapse organization and axon ensheathment (Ank2 and Wrapper) at 14h (Fig. 6F, Supplemental Table S14 and S15). Likewise, all 21 tissue clusters can be examined in our web interface.

Comparing transcriptome and proteome to study translational delay

We compared our embryogenesis proteome with the transcriptome generated as part of the modENCODE project (Graveley et al. 2011). In agreement with the transcriptome analysis, we found that the general protein complexity is increased during embryogenesis (Fig. 7A and Supplemental Fig. S7A). Similar to a previous study in yeast (Fournier et al. 2010), we found only a moderate correlation (maximum R=0.5) between transcriptome and proteome and noted that the best correlation is non-synchronous, showing a 4-5 hours proteome delay (Fig. 7B).

By multidimensional scaling followed by clustering, we sub-grouped the RNA/protein expression profiles into 6 clusters (Fig. 7C, Supplemental Table S16). In the majority of cases, the mRNA is more abundant at early time points, while the protein expression peaks at later stages.

Except for cluster 1, the remaining clusters illustrate different behavior of RNA and protein during embryogenesis. We observed a temporal proteome delay in clusters 5 and 6: while the RNA expression peaks around 7h, proteins steadily upregulate later in embryogenesis, putatively due to translational control mechanisms (Fig. 7C).

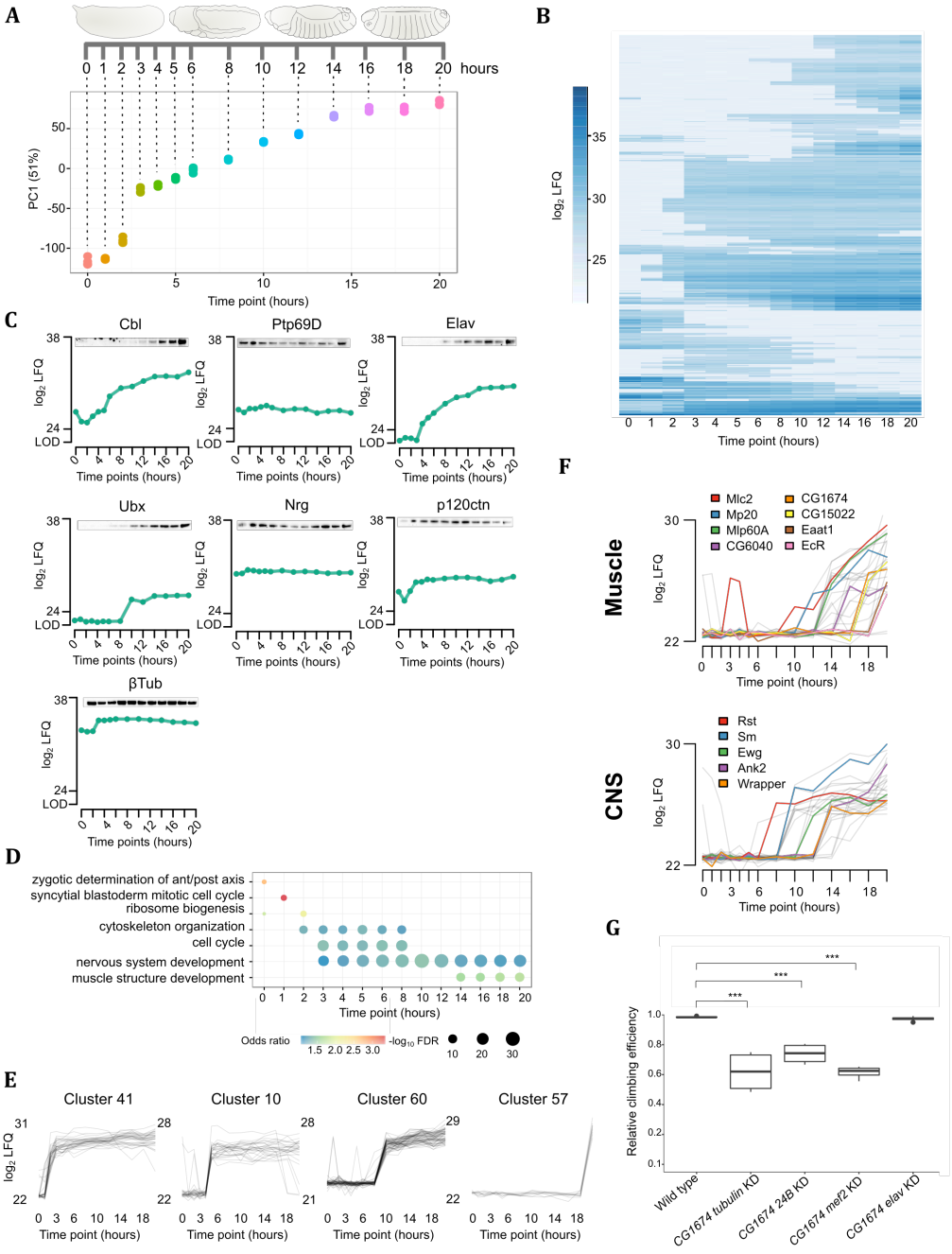


Fig. 6. The embryogenesis proteome time course.
(A) Scheme indicating the collected time points. PCA shows high reproducibility of replicates and the first component shows high correlation with developmental progression (R=0.93). (B) Heat map of log2 LFQ expression values for 1,644 developmentally regulated protein groups in embryogenesis. (C) Western blots of seven selected proteins validate their

temporal expression profile from the proteomics screen. (D) Dot plot connecting the selected enriched GO terms with developmental progression. The circle size indicates the odds ratio of each GO term category. (E) The regulated protein groups have been assigned automatically to 70 clusters based on expression profiles of which four representative clusters with an upregulation at 2-3h (cluster 41), 5h (cluster 10), 10h (cluster 60) and 20h (cluster 57) are shown. (F) Profiles of tissue-specific protein expression created by integrating RNA fluorescence in situ hybridization data. Muscle and central nervous system (CNS) clusters were chosen as examples. (G) Ubiquitous (tubulin-GAL4) and mesodermal (24B- and mef2-GAL4), but not neuronal (elav-GAL4) knockdown of CG1674 results in reduced locomotion activity (Dunnett's test; *** p-value <0.001).

Quantification of protein isoforms during embryogenesis

As distinct protein isoforms may show differential developmental regulation, we mined our proteomic data for protein isoforms. We found 34 genes with various quantified isoforms, some of them showing differential expression such as lola, mod(mdg)4 and Rtnl1 (Fig. 7D and Supplemental Fig. S7B). We further validated our isoform quantitation by immunoblotting following the expression of Lola- RAA/RI (also known as Lola-K) (Giniger et al. 1994). While Lola-RAA/RI is highly expressed at 20h (Fig. 7E), its mRNA shows an expression peak at 14h shown by in situ hybridization (Fig. 7F). This underscores again the importance of a developmental proteome as an addition to transcriptomic studies.

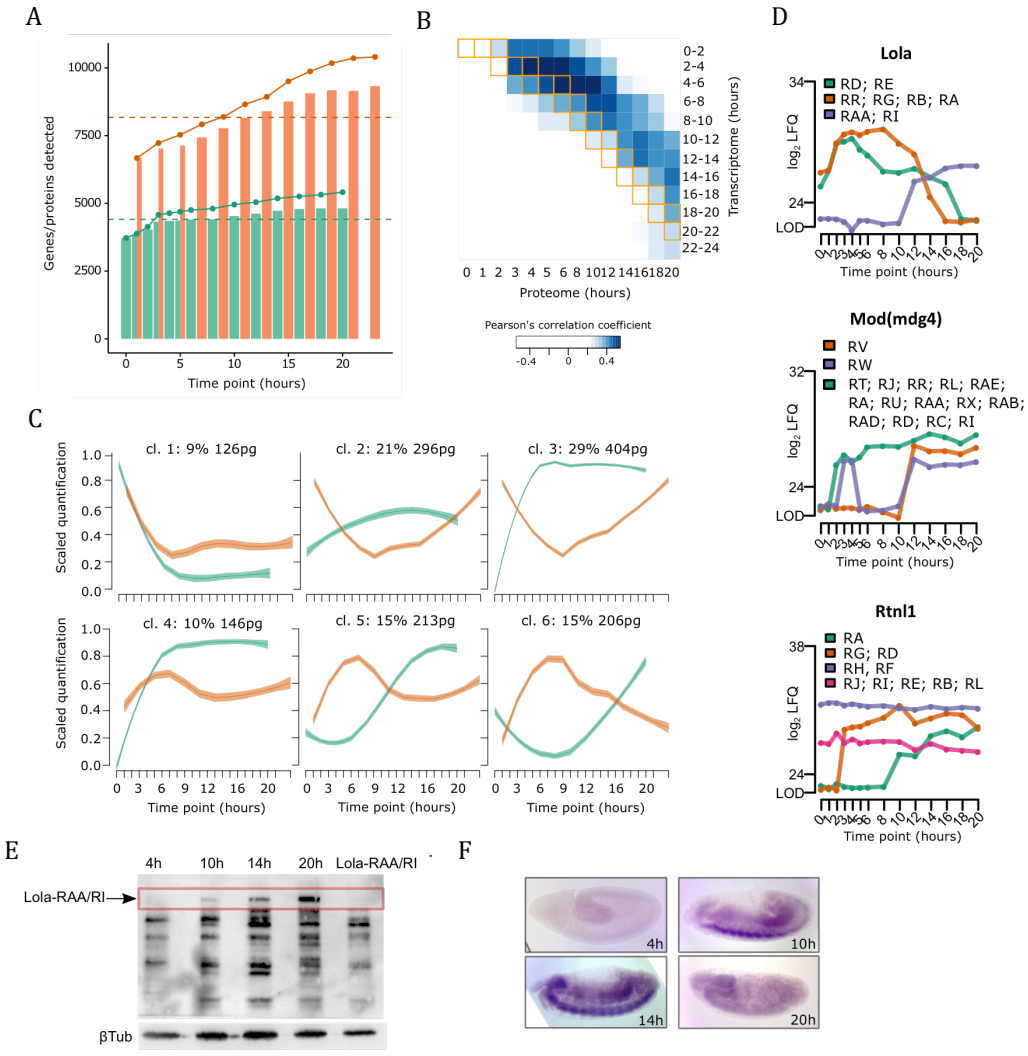


Fig. 7. Temporal transcriptome/proteome dynamics and isoform quantitation.
(A) Plot showing the number (bars) of detected transcripts (orange) and proteins (green) at each time point. The solid line depicts the cumulative sum of unique transcripts (orange) and proteins (green). The dashed line represents the median across all time points. (B) Heat map displaying the Pearson correlation between transcript and protein expression levels. Matching time points between the two datasets are indicated by orange boxes. (C) Median scaled quantification plotted after clustering of the first PCA component of RNA (orange) and protein (green) expression into six different categories. Shaded regions display the standard error of the fitted line. (D) Expression profiles with isoforms-specific information of three proteins: Lola, Mod(mdg4) and Rtnl1. Isoforms are colored according to the legend. (E) Validation of Lola-RAA/Lola-RI isoform quantitation by immunoblotting against Lola at four selected time points. Protein lysate of lola-RAA/lola-RI mutant embryos at 20h were used to identify the isoform-specific band (arrow) corresponding to Lola-RAA/Lola-RI. Beta-tubulin

was used as a loading control. (F) RNA levels were determined by in situ hybridization at the selected time points with a specific probe for lola-RAA/lola-RI.

DISCUSSION

We generated high quality proteome data sets for embryogenesis and the full life cycle of Drosophila melanogaster that close the gap for systematic developmental investigation of protein expression. Both proteomes cover nearly 8,000 and 5,500 protein groups during life cycle and embryogenesis, respectively, accounting for at least one third of annotated Drosophila genes. However, while these two data sets are larger than previous ones, they are not complete. Especially low abundant proteins or proteins that are highly expressed in a restricted number of small tissues will likely not be present in our proteomes. Thus, a not quantified protein can either be absent in this stage or expressed below our limit of detection (LOD) enforced by the mass spectrometry measurement. Nevertheless, these large-scale data sets allow us to assess the developmental expression of proteins and protein isoforms, report maternally provided proteins, validate small proteins (≤100 amino acids), identify Cyp9f3psi as an expressed protein-coding gene and describe peptides originating from non-coding regions.

We scored significantly developmental-regulated protein groups: 1,535 for the whole life cycle and 1,644 for embryogenesis. Nearly half of them are not in-depth characterized suggesting a large area of developmental gene regulation still to be discovered.

We used our data to follow the well-characterized regulation by the hormone ecdysone at a protein level. This revealed intriguing differences to previously reported transcriptome analysis (Beckstead et al. 2005; Gonsalves et al. 2011). For several ecdysone-induced genes, the protein abundance relies on protein stability rather than the presence of RNA transcripts. Overall, transcript abundance and protein levels correlate only modestly. The same observation holds true even considering the temporal delay between transcript and protein expression. The temporal difference in RNA and protein expression needs to be

taken into account when studying phenotypic differences of protein-coding genes using mRNA as a proxy.

As previous transcriptomic studies reported maternally loaded RNAs, our proteomic data enables systematic identification of maternally provided proteins. Here we catalogue not yet reported maternally loaded proteins such as CG14309, CG14834 and CG12398, whose functions in early development need further investigations. For instance, the knockdown of the maternally loaded protein CG17018 results in a severe defect in embryo development.

To gain further insights, we complemented our data sets with other available published data. For example, to deconvolute tissue-specific expression information, we merged our embryogenesis proteome to RNA in situ hybridization data (Lécuyer et al. 2007). This allowed us to pinpoint individual proteins showing tissue-specific developmental regulation, as exemplified with the impaired muscular phenotypes of CG1674 and CG1640 knockdown lines. Additionally, this analysis can be extended to other tissues to uncover currently unknown proteins, which might play an essential role in the development of a specific tissue. This underscores the power to combine available high-quality Drosophila data sets to achieve a more holistic model for developmental gene regulation. While we highlight several interesting results, the entire data sets are available at http://www.butterlab.org/flydev.

METHODS

Collection of embryos, larvae, pupae and adult flies

Population cages of wild type Oregon R flies containing only fertilized females were maintained at 25°C. For the whole life cycle comparative analysis, embryos were collected from cages on agar apple juice plates in 2 hour laying time windows and processed immediately (0-2h) or aged at 25°C for the required time (4-6h, 10-12h, 18-20h). Early larval collections were performed from embryo plates whereas crawling larvae and pupae stages were collected directly from flasks at indicated time points. Virgin young flies within 4h after eclosure were collected separately for males and females, as well as one week

old flies (adult flies). For the time course analysis, embryos were collected on apple juice agar plates in 30 min laying time windows, processed immediately (0h time point) or aged at 25°C for the required time. All samples were mechanically lysed prior to mass spectrometry sample preparation (see supplemental materials for detailed descriptions).

Mass spectrometry measurement and label-free analysis

Peptides were separated by nanoflow liquid chromatography on an EASY-nLC 1000 system (Thermo) coupled to a Q Exactive Plus mass spectrometer (Thermo). Separation was achieved by a 25 cm capillary (New Objective) packed in-house with ReproSil-Pur C18-AQ 1.9 µm resin (Dr. Maisch). Peptides were separated chromatographically by a 280 min gradient from 2% to 40% acetonitrile in 0.5% formic acid with a flow rate of 200 nl/min. Spray voltage was set between 2.4-2.6 kV. The instrument was operated in data-dependent mode (DDA) performing a top15 MS/MS per MS full scan. Isotope patterns with unassigned and charge state 1 were excluded. MS scans were conducted with 70,000 and MS/MS scans with 17,500 resolution. The raw measurement files were analyzed with MaxQuant 1.5.2.8 standard settings except LFQ (Cox et al., 2014) and match between run options were activated as well as quantitation was performed on unique peptides only. The raw data was searched against the translated ENSEMBL transcript databases (release 79) of D. melanogaster (30,362 translated entries) and the S. cerevisiae protein database (6,692 entries). Known contaminants, protein groups only identified by site and reverse hits of the MaxQuant results were removed. In the life cycle data set, the imputation was performed in two distinct ways for proteins with a measured intensity (raw) missing a LFQ intensity or proteins with no intensity value. In the first case, values were imputed from a normal distribution with a mean value shifted by -0.6 from the mean value of all measured LFQ intensities and half of the standard deviation. In contrast, proteins with no intensity value were replaced with the smallest measured value in the set. For the embryo time course, missing values were drawn from a distribution calculated with the

logspline R package (Kooperberg 2016). For cases where 3 or more replicates were measured, the mean of the measured replicates was used as the mean parameter of the distribution. Otherwise, the average of the two neighboring time points was used. In cases of no measured values in neighboring time points, or for proteins measured only in 1 replicate with no surrounding values, a fixed value of 22.5 close to the LOD was used.

Bioinformatics analysis

Significant changes during the lifecycle were calculated by analysis of variance (ANOVA), flagging stage-specific proteins those with FDR<0.01 (Benjamini-Hochberg-procedure) and present in either one unique stage or differing in one stage compared to the rest (log2 LFQ FC>4 in all stages, allowing only one not fulfilling the condition). The effect of the differences was assessed calculating Cohen's effect size and the Tukey HSD post-hoc test. The Gini ratio was used to measure the stability of protein abundance throughout time. Automatic clustering of genes and samples was performed using Affinity Propagation (Frey et al 2007) on the significant proteins, taking negative squared Euclidean distances as a measure of similarity. The goodness of the clusters was assessed from the Silhouette information according to the given clustering. Gene set enrichment analysis (GSEA) was done in R (R core team 2017), followed by a strategy of scoring similar (redundant) terms calculating the information content (IC) between two terms. Results were presented as a treemap or a scatterplot of terms clustered based on the first 2 components of a PCA of the IC similarity scores. For the embryo development significant changes of protein abundance along the time course was assessed. FPKM levels for FlyBase 5.12 Transcripts from short poly(A)+ RNA-seq (Graveley et al. 2011) and localization data from http://fly-fish.ccbr.utoronto.ca were integrated with our proteome data.

DATA ACCESS

The mass spectrometry raw data from this study have been submitted to the ProteomeXchange (http://www.proteomexchange.org) under the data set identifier PXD005691 (life cycle) and PXD005713 (embryogenesis).

DISCLOSURE DECLARATION

The authors declare that there is no conflict of interest.

ACKNOWLEDGEMENT

We thank ... for training as well as ..., ... and ... for critical reading of the manuscript. Antibodies were obtained from the Developmental Studies Hybridoma Bank, created by the NICHD of the NIH or kindly provided by .... We acknowledge support of the Zentrum für Datenverarbeitung (ZDV) at the University of Mainz in hosting the web application. The study was partly funded by the Rhineland Palatinate Forschungsschwerpunkt GeneRED (Gene Regulation in Evolution and Development). DK is supported by the National Research Foundation Singapore and the Singapore Ministry of Education under its Research Centres of Excellence initiative.

REFERENCES

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al. 2000. The genome sequence of Drosophila melanogaster, Science 287: 2185–2195.

Amos LA. 2008. The tektin family of microtubule-stabilizing proteins, Genome Biol 9: 229.

Aspden JL, Eyre-Walker YC, Phillips RJ, Amin U, Mumtaz MAS, Brocard M, Couso J-P. 2014. Extensive translation of small Open Reading Frames revealed by Poly-Ribo-Seq, Elife 3: e03528.

Bahadorani S, Hilliker AJ. 2008. Antioxidants cannot suppress the lethal phenotype of a Drosophila melanogaster model of Huntington's disease, Genome 51: 392–395.

Beckstead RB, Lam G, Thummel CS. 2005. The genomic response to 20-hydroxyecdysone at the onset of Drosophila metamorphosis, Genome Biol 6: R99.

Bonaldi T, Straub T, Cox J, Kumar C, Becker PB, Mann M. 2008. Combined use of RNAi and quantitative proteomics to study gene function in Drosophila, Mol Cell 31: 762–772.

Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, Booth BW, Wen J, Park S, Suzuki AM, et al. 2014. Diversity and dynamics of the Drosophila transcriptome, Nature 512: 393–399.

Brunner E, Ahrens CH, Mohanty S, Baetschmann H, Loevenich S, Potthast F, Deutsch EW, Panse C, Lichtenberg U de, Rinner O, et al. 2007. A high-quality catalog of the Drosophila melanogaster proteome, Nat Biotechnol 25: 576–583.

Butter F, Bucerius F, Michel M, Cicova Z, Mann M, Janzen CJ. 2013. Comparative proteomics of two life cycle stages of stable isotope-labeled Trypanosoma brucei reveals novel components of the parasite's host adaptation machinery, Mol Cell Proteomics 12: 172–179.

Campos-Ortega JA, Hartenstein V. 1997. The embryonic development of Drosophila melanogaster, 2nd edn. Springer, Berlin, London.

Chang YC, Tang HW, Liang SY, Pu TH, Meng TC, Khoo KH, Chen GC. 2013. Evaluation of Drosophila metabolic labeling strategies for in vivo quantitative proteomic analyses with applications to early pupa formation and amino acid starvation, J Proteome Res 3: 2138-2150.

Chao YC, Donahue KM, Pokrywka NJ, Stephenson EC. 1991. Sequence of swallow, a gene required for the localization of bicoid message in Drosophila eggs, Dev Genet 12: 333–341.

Chintapalli VR, Wang J, Dow JAT. 2007. Using FlyAtlas to identify better Drosophila melanogaster models of human disease, Nat Genet 39: 715–720.

Cox J, Hein MY, Luber CA, Paron I, Nagaraj N, Mann M. 2014. Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ, Mol Cell Proteomics 13: 2513–2526.

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.

Dejung M, Subota I, Bucerius F, Dindar G, Freiwald A, Engstler M, Boshart M, Butter F, Janzen CJ. 2016. Quantitative Proteomics Uncovers Novel Factors Involved in Developmental Differentiation of Trypanosoma brucei, PLoS Pathog 12: e1005439.

Dorus S, Busby SA, Gerike U, Shabanowitz J, Hunt DF, Karr TL. 2006. Genomic and functional evolution of the Drosophila melanogaster sperm proteome, Nat Genet 38: 1440–1445.

Edgar BA, Datar SA. 1996. Zygotic degradation of two maternal Cdc25 mRNAs terminates Drosophila's early cell cycle program, Genes Dev 10: 1966–1977.

Fabre B, Korona D, Groen A, Vowinckel J, Gatto L, Deery MJ, Ralser M, Russell S, Lilley KS. 2016. nalysis of Drosophila melanogaster proteome dynamics during embryonic development by a combination of label-free proteomics approaches, Proteomics 16: 2068-2080.

Fakhouri M, Elalayli M, Sherling D, Hall JD, Miller E, Sun X, Wells L, LeMosy EK. 2006. Minor proteins and enzymes of the Drosophila eggshell matrix, Dev Biol 293: 127–141.

Fournier ML, Paulson A, Pavelka N, Mosley AL, Gaudenz K, Bradford WD, Glynn E, Li H, Sardiu ME, Fleharty B, et al. 2010. Delayed correlation of mRNA and protein expression in rapamycin-treated cells and a role for Ggc1 in cellular sensitivity to rapamycin, Mol Cell Proteomics 9: 271–284.

Frey BJ, Dueck D. 2007. Clustering by passing messages between data points. Science 315: 972-976.

Giniger E, Tietje K, Jan LY, Jan YN. 1994. lola encodes a putative transcription factor required for axon growth and guidance in Drosophila, Development 120: 1385–1398.

Gonsalves SE, Neal SJ, Kehoe AS, Westwood JT. 2011. Genome-wide examination of the transcriptional response to ecdysteroids 20-hydroxyecdysone and ponasterone A in Drosophila melanogaster, BMC Genomics 12: 475.

Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW, et al. 2011. The developmental transcriptome of Drosophila melanogaster, Nature 471: 473–479.

Griffin TJ, Gygi SP, Ideker T, Rist B, Eng J, Hood L, Aebersold R. 2002. Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae, Mol Cell Proteomics 1: 323–333.

Grün D, Kirchner M, Thierfelder N, Stoeckius M, Selbach M, Rajewsky N. 2014. Conservation of mRNA and protein expression during development of C. elegans, Cell Rep 6: 565–577.

Harrison PM, Milburn D, Zhang Z, Bertone P, Gerstein M. 2003. Identification of pseudogenes in the Drosophila melanogaster genome, Nucleic Acids Res 31: 1033–1037.

Jambor H, Surendranath V, Kalinka AT, Mejstrik P, Saalfeld S, Tomancak P. 2015. Systematic imaging reveals features and changing localization of mRNAs in Drosophila development, Elife 4.

Kalinka AT, Varga KM, Gerrard DT, Preibisch S, Corcoran DL, Jarrells J, Ohler U, Bergman CM, Tomancak P. 2010. Gene expression divergence recapitulates the developmental hourglass model, Nature 468: 811–814.

Kappei D, Butter F, Benda C, Scheibe M, Draskovic I, Stevense M, Novo CL, Basquin C, Araki M, Araki K, et al. 2013. HOT1 is a mammalian direct telomere repeat-binding protein contributing to telomerase recruitment, EMBO J 32: 1681–1701.

Kharchenko PV, Alekseyenko AA, Schwartz YB, Minoda A, Riddle NC, Ernst J, Sabo PJ, Larschan E, Gorchakov AA, Gu T, et al. 2011. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster, Nature 471: 480–485.

Kondo S, Ueda R. 2013. Highly improved gene targeting by germline-specific Cas9 expression in Drosophila, Genetics 195: 715–721.

Kooperberg. 2016 logspline: Logspline Density Estimation Routines. https://CRAN.R-project.org/package=logspline

Kronja I, Whitfield ZJ, Yuan B, Dzeyk K, Kirkpatrick J, Krijgsveld J, Orr-Weaver TL. 2014. Quantitative proteomics reveals the dynamics of protein changes during Drosophila oocyte maturation and the oocyte-to-embryo transition, Proc Natl Acad Sci U S A 111: 16023–16028.

Ladoukakis E, Pereira V, Magny EG, Eyre-Walker A, Couso JP. 2011. Hundreds of putatively functional small open reading frames in Drosophila, Genome Biol 12: R118.

Lawrence PA. 1992. The making of a fly. The genetics of animal design / Peter A. Lawrence. Blackwell Scientific, Oxford.

Lécuyer E, Yoshida H, Parthasarathy N, Alm C, Babak T, Cerovina T, Hughes TR, Tomancak P, Krause HM. 2007. Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function, Cell 131: 174–187.

Liu N, Lasko P. 2015. Analysis of RNA Interference Lines Identifies New Functions of Maternally-Expressed Genes Involved in Embryonic Patterning in Drosophila melanogaster, G3 (Bethesda) 5: 1025–1034.

Liu Y, Beyer A, Aebersold R. 2016. On the Dependency of Cellular Protein Levels on mRNA Abundance, Cell 165: 535–550.

Luschnig S, Moussian B, Krauss J, Desjeux I, Perkovic J, Nüsslein-Volhard C. 2004. An F1 genetic screen for maternal-effect mutations affecting embryonic pattern formation in Drosophila melanogaster, Genetics 167: 325–342.

Magny EG, Pueyo JI, Pearl FMG, Cespedes MA, Niven JE, Bishop SA, Couso JP. 2013. Conserved regulation of cardiac calcium uptake by peptides encoded in small open reading frames, Science 341: 1116–1120.

Mani SR, Megosh H, Lin H. 2014. PIWI proteins are essential for early Drosophila embryogenesis, Dev Biol 385: 340–349.

Margolin W. 2012. The price of tags in protein localization studies, J Bacteriol 194: 6369–6371.

Matthews BB, Dos Santos G, Crosby MA, Emmert DB, St Pierre SE, Gramates LS, Zhou P, Schroeder AJ, Falls K, Strelets V, et al. 2015. Gene Model Annotations for Drosophila melanogaster: Impact of High-Throughput Data, G3 (Bethesda) 5: 1721–1736.

McGraw LA, Clark AG, Wolfner MF. 2008. Post-mating gene expression profiles of female Drosophila melanogaster in response to time and to four male accessory gland proteins, Genetics 179: 1395–1408.

Morin X, Daneman R, Zavortink M, Chia W. 2001. A protein trap strategy to detect GFP-tagged proteins expressed from their endogenous loci in Drosophila, Proc Natl Acad Sci U S A 98: 15050–15055.

Nagarkar-Jaiswal S, Lee P-T, Campbell ME, Chen K, Anguiano-Zarate S, Gutierrez MC, Busby T, Lin W-W, He Y, Schulze KL, et al. 2015. A library of MiMICs allows tagging of genes and reversible, spatial and temporal knockdown of proteins in Drosophila, Elife 4.

Peshkin L, Wühr M, Pearl E, Haas W, Freeman RM, Gerhart JC, Klein AM, Horb M, Gygi SP, Kirschner MW. 2015. On the Relationship of Protein and mRNA Dynamics in Vertebrate Embryonic Development, Dev Cell 35: 383–394.

R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Ramamurthi KS, Storz G. 2014. The small protein floodgates are opening; now the functional analysis begins, BMC Biol 12: 96.

Sandmann T, Jensen LJ, Jakobsen JS, Karzynski MM, Eichenlaub MP, Bork P, Furlong EEM. 2006. A temporal map of transcription factor activity: mef2 directly regulates target genes at all stages of muscle development, Dev Cell 10: 797–807.

Sarov M, Barz C, Jambor H, Hein MY, Schmied C, Suchold D, Stender B, Janosch S, Kj VV, Krishnan RT, et al. 2016. A genome-wide resource for the analysis of protein localisation in Drosophila, Elife 5.

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: 337–342.

Simmons MJ, Thorp MW, Buschette JT, Peterson K, Cross EW, Bjorklund EL. 2010. Maternal impairment of transposon regulation in Drosophila melanogaster by mutations in the genes aubergine, piwi and Suppressor of variegation 205, Genet Res (Camb) 92: 261–272.

Sowell RA, Hersberger KE, Kaufman TC, Clemmer DE. 2007. Examining the proteome of Drosophila across organism lifespan, J Proteome Res 6: 3637–3647.

Sury MD, Chen J-X, Selbach M. 2010. The SILAC fly allows for accurate protein quantification in vivo, Mol Cell Proteomics 9: 2173–2183.

Tadros W, Lipshitz HD. 2005. Setting the stage for development: mRNA translation and stability during oocyte maturation and egg activation in Drosophila, Dev Dyn 232: 593–608.

Takemori N, Yamamoto M-T. 2009. Proteome mapping of the Drosophila melanogaster male reproductive system, Proteomics 9: 2484–2493.

The modENCODE consortium, Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, Landolin JM, Bristow CA, Ma L, Lin MF, et al. 2010. Identification of functional elements and regulatory circuits by Drosophila modENCODE, Science 330: 1787–1797.

Tomancak P, Berman BP, Beaton A, Weiszmann R, Kwan E, Hartenstein V, Celniker SE, Rubin GM. 2007. Global analysis of patterns of gene expression during Drosophila embryogenesis, Genome Biol 8: R145.

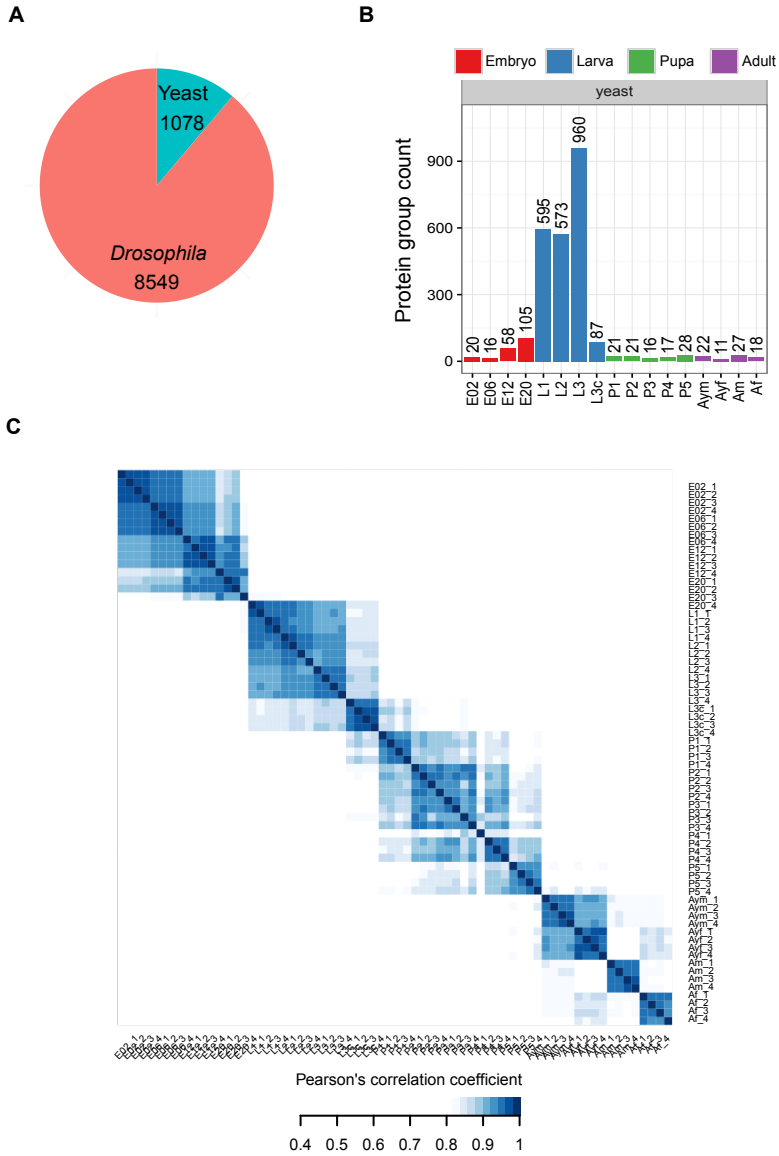
Vogel C, Marcotte EM. 2012. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses, Nat Rev Genet 13: 227–232.

Wasbrough ER, Dorus S, Hester S, Howard-Murkin J, Lilley K, Wilkin E, Polpitiya A, Petritis K, Karr TL. 2010. The Drosophila melanogaster sperm proteome-II (DmSP-II), J Proteomics 73: 2171–2185.

Xing X, Zhang C, Li N, Zhai L, Zhu Y, Yang X, Xu P. 2014. Qualitative and quantitative analysis of the adult Drosophila melanogaster proteome, Proteomics 14: 286–290.

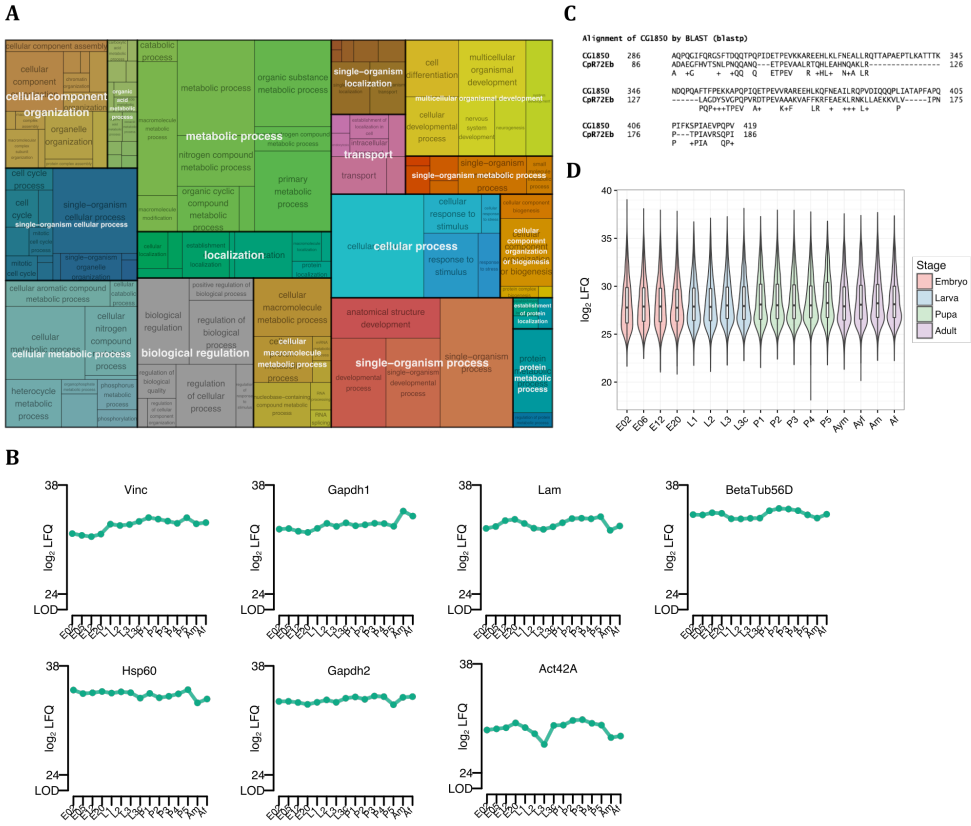
Yamanaka N, Rewitz KF, O'Connor MB. 2013. Ecdysone control of developmental transitions: lessons from Drosophila research, Annu Rev Entomol 58: 497–516.

SUPPLEMENTARY FIGURES



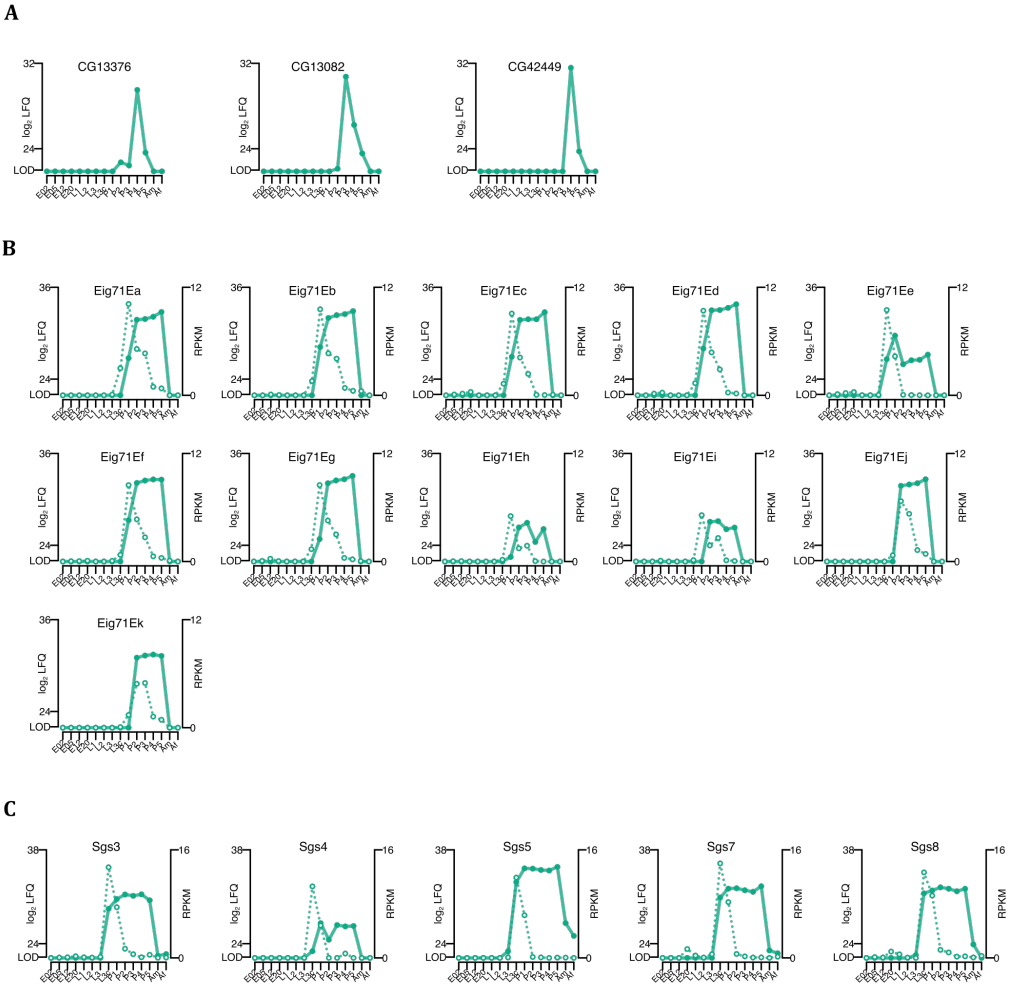
Supplemental Fig. S1:

(A) Pie chart of detected yeast and *Drosophila* protein groups. Yeast proteins represent 12 percent (1078) of the identified proteins in the screen (9627). (B) Nearly all yeast proteins are detected in feeding L1, L2 and early L3 larval stages. (C) Correlation plot for the 17 samples of the life cycle experiment with all four replicates.



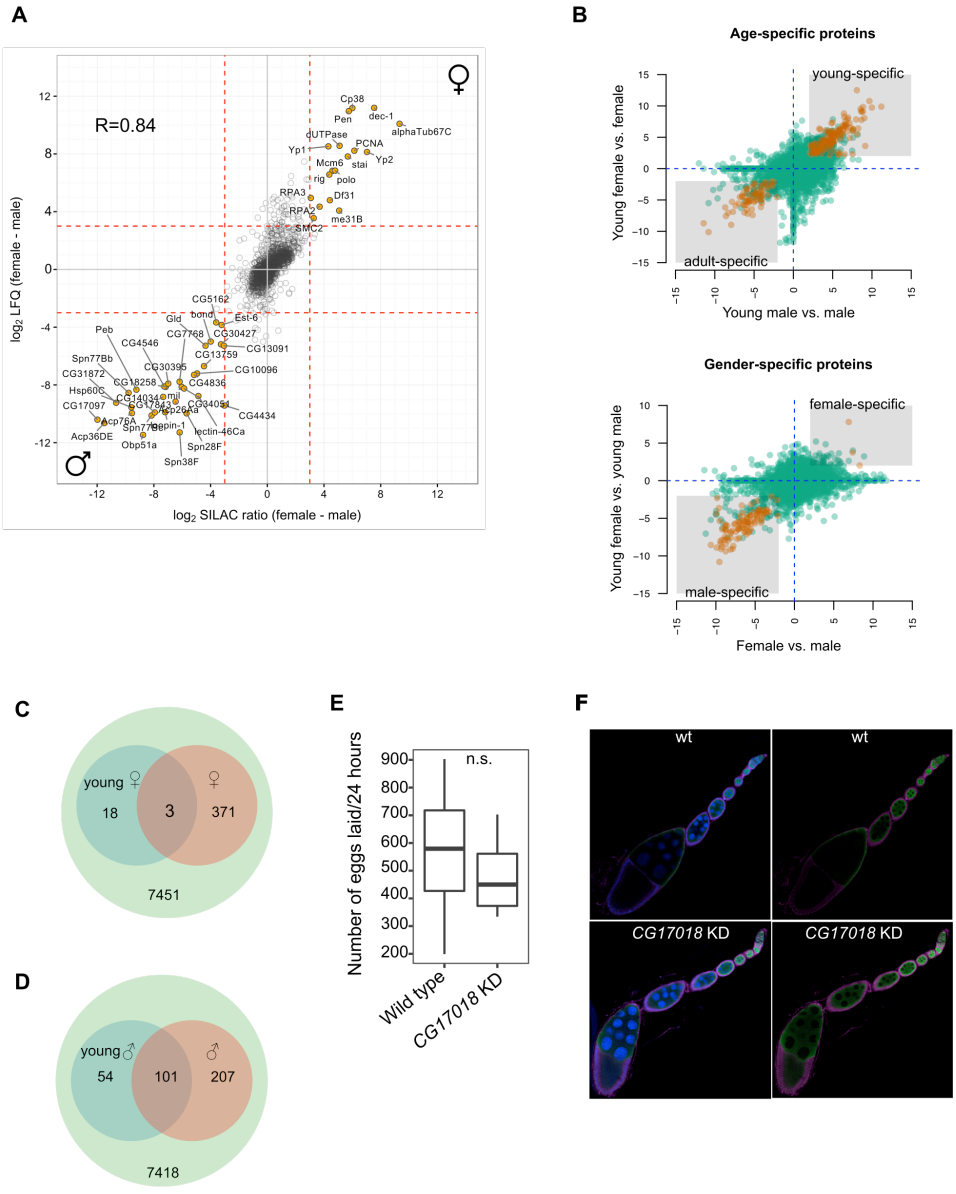
Supplemental Fig. S2:

(A) Treemap with overrepresented GO terms. The area is proportional to the size of the GO terms. (B) Alignment using BLAST between CpR72Eb and CG1850 shows a stretch of moderate sequence similarity between both proteins. (C) Overall protein expression values based on LFQ quantitation shows similar abundance levels and distribution across all conditions of the experiment. (D) Expression profiles of stably and dynamically expressed proteins annotated in Figure 2C and not displayed in Figure 2D.



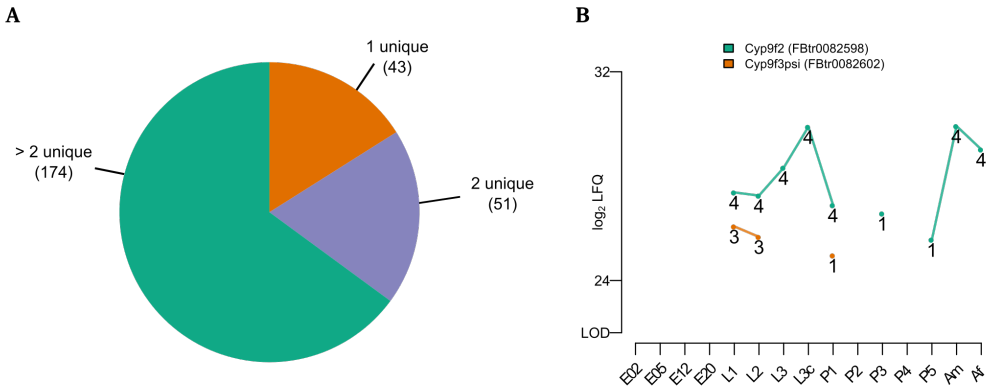
Supplemental Fig. S3:

Expression profiles of uncharacterized proteins upregulating at a single pupal stage (A). RNA (dotted line) and protein (solid line) expression profiles of all quantified Eig72E family proteins (B) and Sgs proteins (C).



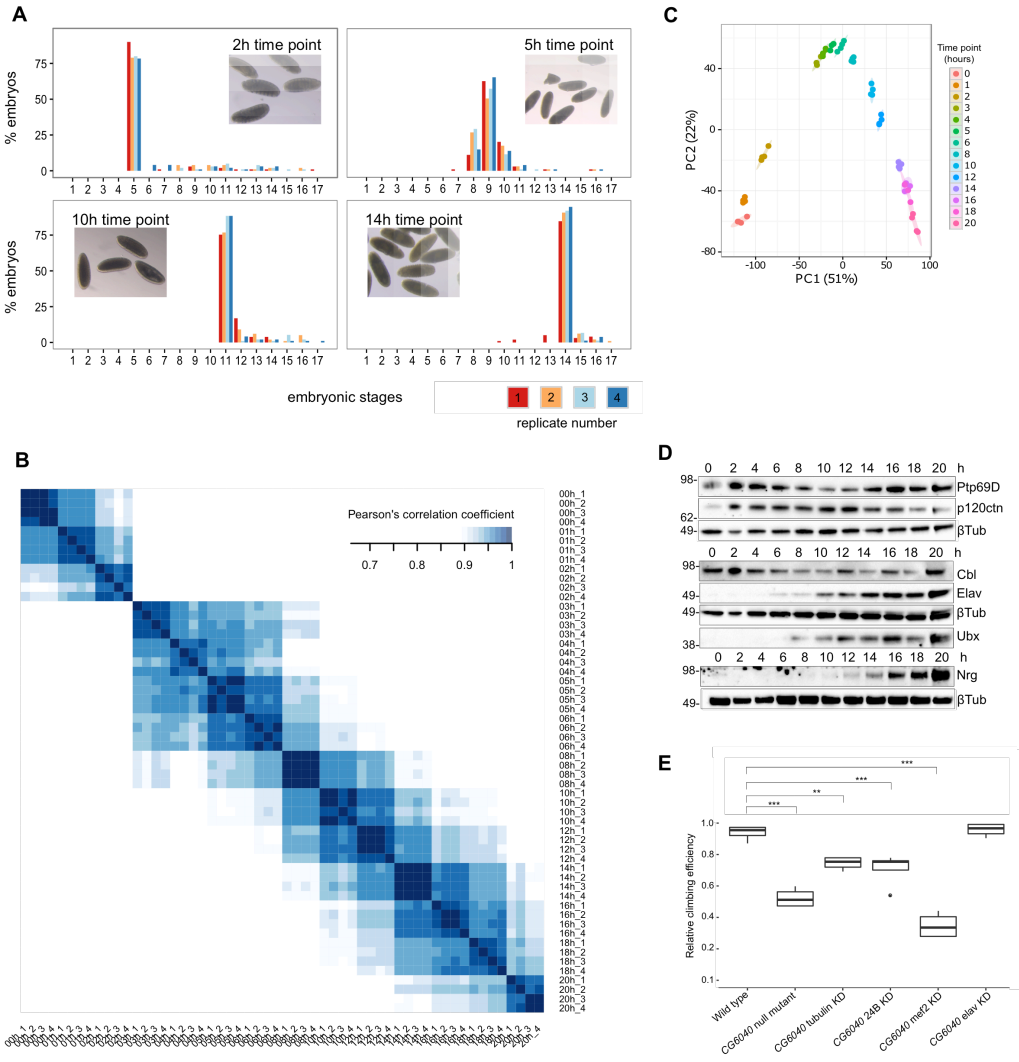
Supplemental Fig. S4:

(A) Comparison of w1118 SILAC data (Sury et al., 2010) to our Oregon R strain LFQ data. More than 8-fold enriched protein groups are highlighted (orange fill) and annotated. (B) Scatter plot showing age-specific and gender-specific proteins. (C) Venn diagram to compare female-specific proteins already present in the young fly. (D) Venn diagram to compare male-specific proteins already present in the young fly. (E) Number of laid eggs from wild type and the *CG17018* knockdown line. (F) Dissected ovaries from 5 day old female flies stained for DAPI (blue), Vasa (green) and 1B1 (purple).



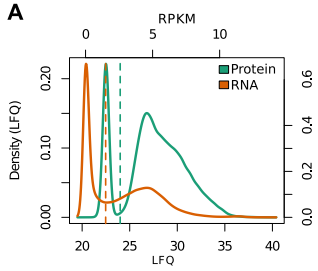
Supplemental Fig. S5:

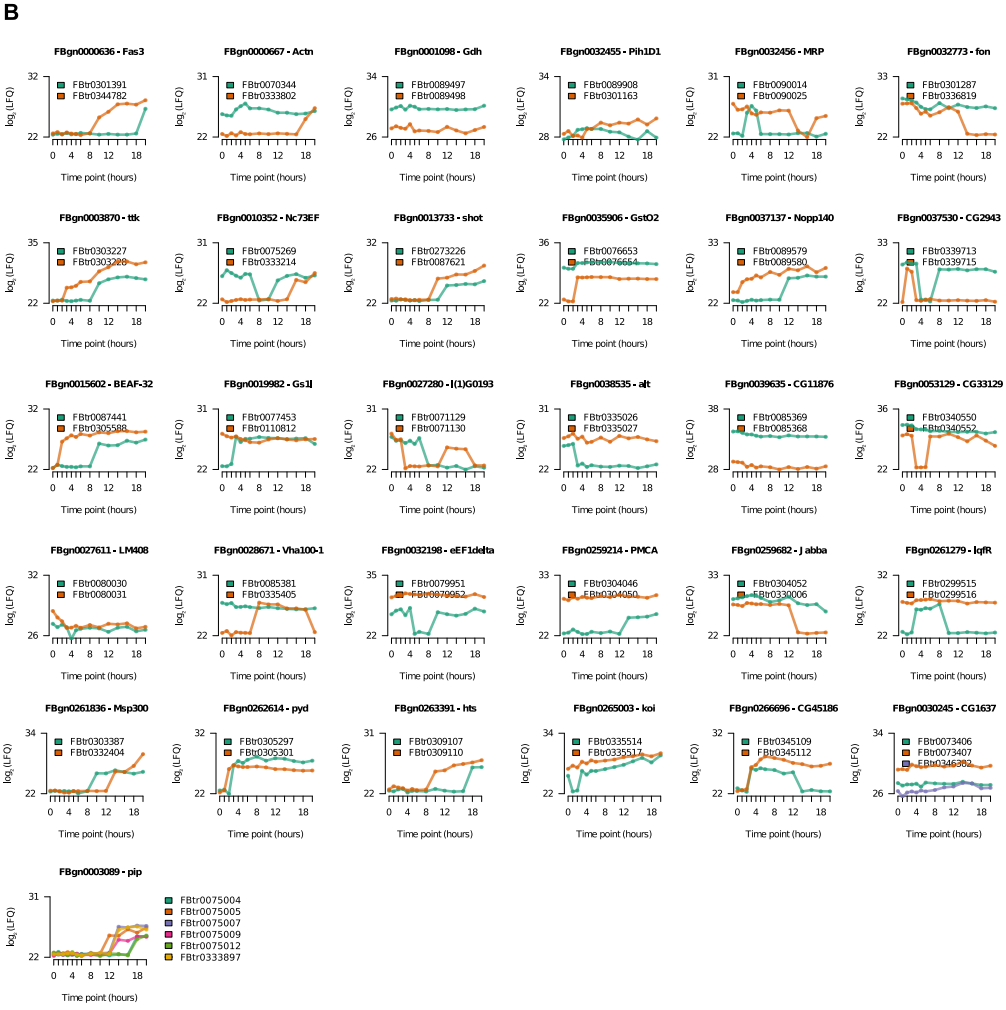
(A) Pie chart showing the distribution of small proteins identified by the number of unique peptides. For a large majority (>80 percent) of the small proteome more than a single unique peptide was measured. (B) Developmental expression profile for the expressed previously annotated pseudogene Cyp9f3Psi in comparison to its paralog Cyp9f2.



Supplemental Fig. S6:

(A) Histogram showing the distribution of embryo stages in the four replicates of selected staged time points with a representative picture. (B) Correlation analysis of the 500 most variant proteins show high correlation between individual biological replicates. (C) PCA analysis with the first two components. Replicates are shown in identical colors and the standard error is represented by elliptic areas. (D) Westerm blot validation of chosen protein profiles in embryogenesis shown in Fig. 6. The respective tubulin loading controls are shown. (E) Ubiquitous (tubulin driver) and mesodermal (24B and mef2 driver), but not neuronal (elav) knockdown of *CG6040* results in reduced locomotion activity (Dunnett's test; ** p-value <0.005, *** p-value < 0.001). (F) *in situ* hybridization of *CG6040* and *CG1674* showing expression in muscle tissue.





Supplemental Fig. 7:

(A) Density plot of transcriptome (RPKM) and proteome (LFQ) quantitation shows bimodal distribution. The cut-offs are chosen to separate the two maxima (dashed lines). (B) Profile plots of the remaining 31 genes (not shown in Figure 7D) with isoform quantitation on protein level during the time course experiment.

SUPPLEMENTAL METHODS

Collection of embryos, larvae, pupae and adult flies

After collection, all embryo samples were dechorionated using 7.5% hypochlorite for 2 min and rinsed with water. For each time point, approximately 20 µl embryo pellets were transferred to PBS buffer for lysis and mass spectrometry measurement. To assess the homogeneity of embryonic stages of each collection, approximately 10% of the sample was fixed and staged. Unused samples were snap-frozen in liquid nitrogen and stored at -80°C. Early larval collections were performed from agar apple juice plates in 2 hour laying time windows. Crawling larvae and pupae stages were collected from flasks at respective time points and rinsed with water.

Fixation and antibody staining for embryo staging

A small fraction (approx. 10%) of each embryo collection during the time course was fixed in fixation buffer (450 µl PBS, 600 µl heptane, 70 µl 37% formaldehyde) for 20 min while agitating. To remove the vitelline membrane, heptane was exchanged for methanol and embryos were vortexed for 2 minutes. Embryos were washed several times in methanol and finally snap-frozen in liquid nitrogen. Fixed embryos were rinsed three times in PBT (0.1% Triton-X-100 in PBS) for 10 min and incubated with the 4D9-engrailed/invected-s antibody (1:7 in PBT, Developmental Studies Hybridoma Bank (DSHB)) at 4°C while slowly agitating. Embryos were washed three times in PBT for 10 min and incubated with the anti-mouse antibody conjugated with alkaline phosphatase (AP) (1:250 in PBT, Jackson ImmunoResearch) for 2 hours at RT. Three washes in PBT were followed by a 5 min wash in AP detection buffer (0.1 M NaCl, 0.05 M MgCl2, 0.1 M Tris pH 9.5, 0.1% Tween20). The AP staining solution (150 µg/mL nitro blue tetrazolium (NBT) and 75 µg/mL 5-bromo-4-chloro-3-indolyl-phosphate (BCIP) in AP buffer) was added, embryos were transferred to a small dish and the color reaction was monitored using a binocular. To stop the AP reaction, embryos were rinsed three times in PBT and incubated for 10 min in methanol. After three PBT and three PBS

washes, embryos were stored in 1 mL 70% glycerol at RT. Staging was done according to morphology and antibody staining.

Mass spectrometry sample preparation

For proteome analysis of the whole life cycle, snap-frozen samples were mechanically lysed in lysis buffer (140 mM NaCl, 10 mM Tris-HCl, pH 8.0, 0.5 mM EDTA and 1x protease inhibitor (Roche)) by bead milling using 0.5 mm diameter zircona/silica beads (Carl-Roth). Bead milling was done three times for 30 sec at 6800 rpm at 4°C in a tissue lyzer (Precellys). Homogenates were collected by centrifugation and proteins were acetone precipitated. Protein pellets were resuspended in 1x NuPAGE sample loading buffer (1x LDS) supplemented with 0.1 M DTT, boiled for 10 min at 95°C, sonicated for 10 min and proteins were separated on a 4-12% NuPAGE Bis/Tris gel for 10 min at 180 V in MOPS buffer. For the embryonic time course analysis, embryos in PBS were homogenized with a microtube pestle, cells were pelleted at 1000 x g for 5 min at 4°C and resuspended in 1x LDS buffer complemented with 0.1 M DTT. Samples were boiled for 10 min at 80°C and proteins were separated on a 4-12% NuPAGE Bis/Tris gel for 10 min at 180 V in MOPS buffer. In-gel digestion and MS analysis was done as essentially described (Kappei et al. 2013).

Western blotting

100 embryos were homogenized in 50 µl of lysis buffer (140 mM NaCl, 10 mM Tris-HCl pH 8, 1 mM EDTA pH 8, 0.5% Triton X-100) using a microtube pestle (8-10 strokes). After 30 min incubation at 4°C on a rotation wheel, the lysate was centrifuged at maximum speed for 20 min and the supernatant transferred to a new tube. The previous two steps were repeated four times until the lysate was clear. Total embryonic protein extract was separated by SDS-PAGE. Western blot analysis with affinity purified anti-Lola antibody (1:500, kindly provided by Edward Giniger) was performed by standard methods and visualized using ultra-sensitive enhanced chemiluminescent reagent (Thermo). Anti-β-Tubulin antibody (Covance catalog #MMS-410P, BioLegend) was used at

a 1:2000 dilution. Additionally, Lola-RAA/RI depleted flies were generated as previously described (Kondo and Ueda 2013) using a gRNA (GTGTTGCACGTAAAGAAGCT) in exon 21 leading to a 2-bp deletion (Chr2R:10510060-10510061).

Embryo samples of selected time points (0h, 2h, 4h, 6h, 8h, 10h, 12h, 14h, 16h, 18h, 20h) prepared for mass spectrometry were used for western blot validation of protein expression profiles during embryogenesis. Proteins were separated by SDS-PAGE and blotted onto a nitrocellulose membrane, and subsequently blocked with 5% milk powder in TBST (0.1% Tween20 in 1xTBS). The following antibodies from DSHB were used in a 1:50 dilution (5% milk powder TBST): anti-p120ctn (p1B2), anti-Nrg (BP 102), anti-Ubx (kindly provided by Christian Berger), anti-PTP69D (3F11), anti-Cbl (8C4). Anti-Elav (7E8A10, DSHB) raised in rat was used 1:200 in 5% milk powder TBST.

in situ hybridization

Primers were designed to amplify a unique region within respective coding sequences using a reverse primer containing the SP6 sequence. The PCR was performed on embryonic cDNA using Phusion DNA polymerase (NEB). Amplicon size comprised 861 bp for lola-RAA/RI, 1095 bp for CG6040 and 1070 bp for CG1674. 250 ng of template PCR product was used to perform in vitro transcription using the SP6 RNA polymerase (Roche) and DIG labeled UTP (Roche). The reaction was incubated over night at 37°C and probes were carbonated to approximately 500 bp using carbonate buffer. The probes were then ethanol precipitated and resuspended in DEPC treated water to obtain a concentration of 100 ng/µl. Probes were diluted 1:50 in hybridization buffer for in situ hybridization. Embryos were fixed for 25 min in fixation solution (400 µl PBS, 500 µl n-heptane, 100 µl 37% formaldehyde) while shaking at RT. After washing in methanol several times the embryos were snap-frozen in liquid nitrogen and stored at -80°C. Embryos were gradually transferred into PBT (0.1% Tween20 in PBS), followed by three washes for 15 min, and finally into HB4 hybridization buffer (50 ml formamide, 25 ml 20xSSC buffer (3 M NaCl and

0.3 mM trisodium citrate-HCl pH 7.0), 200 µl Heparin (50 mg/ml), 100 µl Tween20, 500 mg Torula Yeast RNA extract). After equilibration at RT, embryos were pre-hybridized in HB4 at 56°C for several hours. Upon denaturation of the diluted RNA probe at 80°C for 10 min, embryos were hybridized over night at 65°C. Embryos were subsequently incubated in washing buffer (formamide:2xSSC (1:1) and 0.1% Tween20) for 30 min at 65°C and transferred into PBT at RT before they were incubated with anti-DIG-AP antibody (1:1000, Roche) for 2h at RT. Upon several washes in PBT and one rinse with AP buffer probes were visualized using the NBT/BCIP solution in AP buffer (1:100). Primers used in 5'-3' orientation:

lola-RAA/RI_fwd	AACCACAACAATTGCCACACATCATC
lola-RAA/RI_rev	GAGAATGGTGTAGCTCTTGCTC
CG6040_fwd	CCTTTGCCGCCTTAAAACTGG
CG6040_rev	CGCTACCCAAGCTAATGCCG
CG1674_fwd	CACTAAAGCAGACCTTGTTTCG
CG1674_ rev	TTTCGCACTGCTGTGAAG

Locomotion assay

Freshly hatched male and female flies of the respective genotype were separated and directly placed into measuring cylinders. The locomotion was assessed using the climbing assay described previously (Bahadorani and Hilliker 2008). Flies were tapped to the bottom and flies passing 8 cm in 10 or 5 seconds, respectively, were counted. Measurements were repeated four times in two independent biological replicates for each of the shRNA expressing-lines.

Fertility assay

20 young female flies were separated upon hatching and mated with 5 males for three days. Upon fertilization, flies were transferred onto fresh agar plates every 24 hours, eggs were counted and allowed to develop for further 36 hours. The hatching rate was determined by counting the number of unhatched eggs for each lay. Counting was done on four subsequent days. Experiments were

performed in three biological replicates with two different RNAi knockdown-lines.

Cuticle preparation

Cuticle preparations were prepared as previously described (Liu and Lasko 2015). Embryos were dechorionated for 2 min with 50% bleach, washed with water, transferred into an Eppendorf tube and washed with PBT (0.1% Tween20 in PBS). The supernatant was removed and Hoyer's medium (30 g gum Arabic, 200 g chloral hydrate, 20 g glycerol ad 50 ml water) was added to cover the embryos. The tube was incubated over night at 65°C, embryos were mounted onto a glass slide and examined under dark field illumination.

Immunohistochemistry

Ovaries were dissected from 5 days-old virgin females in cold PBS and fixed in 5% formaldehyde for 20 min. Samples were washed three times for 10 min with PBT (0.3% Triton X-100 in PBS) and blocked for 20 min in PBT + 5% donkey serum at RT. Samples were incubated with primary antibodies (rabbit anti-Vasa, 1:500; mouse anti-1B1, 1:100) over night at 4°C, washed three times for 10 min with PBT and incubated with secondary antibody for two hours at RT. Samples were mounted in Vectashield and examined using a confocal Leica SP5.

EXTENDED BIOINFORMATICS ANALYSIS

Dynamicity of protein profiles

We used the Gini ratio (Damgaard and Weiner 2000; Gini 1912) to measure the stability of protein abundance throughout time. The Gini ratio calculates a score ranging from 0 to 1, which depicts the normalized mean difference of LFQ values between every possible combination of two stages for each protein:

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n}|x_i - x_j|}{2n^2\mu}$$

where "n" is the number of stages, "xi" the protein quantification (LFQ) at stage "i", and "μ" the average protein quantification throughout time. The minimum score refers to proteins that are stably expressed regardless of their average quantitation, while proteins having high abundance in only one stage present scores close to 1.

Significantly changing proteins throughout the life cycle

We used analysis of variance (ANOVA) in order to detect significant changes in protein expression in the life cycle data set. The resulting p-values were adjusted by Benjamini and Hochberg to control for the false discovery rate (FDR). We defined a protein as differentially expressed with a cut-off of 1% FDR. To identify in which stages proteins significantly change, we used the Tukey HSD post-hoc test. The test defines the Honest Significant Difference as the minimal distance between two groups to be considered statistically significant. To have a measure of the strength of the changes, we also calculated the effect size as suggested for one-way ANOVA analysis (Cohen 1988):

$$f = \sqrt{\frac{\sum_{i=1}^{k} p_i * (\mu_i - \mu)}{\sigma^2}}$$

Stage-specific proteins of the life cycle

To classify proteins into stages, we required that they are differentially expressed at 1% FDR and either detected in only one stage (embryo, larva, pupa or adult) or showing high differences in abundance (LFQ fold change > 4) in at least two other stages.

Getting the significant changes during the embryo development

We used the time-course (Tai 2007) and q-value (Storey et al. 2015) packages to assess significant changes in protein expression during time. The timecourse

package implements a multivariate empirical Bayes method to calculate moderated T^2-statistics from longitudinal data, taking into account replicate information and correlation among gene expression along time points.
The significance of the T^2-statistics was empirically estimated using the q-value package, by comparing the obtained T^2-statistics with the ones obtained from bootstrapping the original data. We performed a 1000-times bootstrap of each protein, permuting with replacement the values and calculating the statistic again. Then, we calculated the empirical p-values comparing the statistics from the original data and the pool of null statistics (bootstrapped data). To control the false discovery rate we use the q-value function of the same package. Based on the distribution of our q-values, we set the significance cut-off at an FDR adjusted p-value of 0.0001.

Clustering strategies

We use Affinity Propagation (Bodenhofer et al. 2011; Frey and Dueck 2007) to cluster the differentially expressed proteins with similar expression profiles into an optimal number of clusters. Affinity Propagation is a well-established method to automatically calculate the best number of fitting clusters to the data. This method takes the data points as potential "exemplars" of clusters and further passing messages between points to decide which are [the best] exemplars and to which exemplar the rest belong to.
We calculated a similarity matrix between protein profiles using the negative Euclidean distance, defined as the negative squared distance between two points:

$$s = -d^r$$

Then we called Affinity Propagation with this similarity matrix, without setting any preference for any protein to be the exemplar of a cluster (default settings are the median of non-infinite values in the similarity matrix), and allowing up to 1000 iterations (i.e. rounds of messages passed between data points).

To reduce the number of clusters of the life cycle data set, we run Affinity Propagation with the preference parameter set at the 10% quantile of the distribution of similarities. For the embryo data set, the default parameters were used instead. The goodness of the clustering was assessed using silhouette plots.

GO analysis

Using the R packages GSEABase (Morgan et al. 2016), GOstats (Falcon and Gentleman 2007) and org.Dm.eg.db (Carlson 2016, date stamp from the source of: 2015-Sep27) we performed gene set enrichment analyses of GO terms (biological processes only).

Due to the hierarchical nature of the GO annotation, usually many terms appear from the same set of genes. To remove redundancy of terms we scored the similarity between terms using the GOSemSim package (Yu et al. 2010). This package implements several methods to calculate the functional similarity of different terms. We used the Relevance method (Schlicker et al. 2006), based on the Information Content of two terms and their closest common ancestor.

Embryogenesis data set: For each cluster of the embryo data set we performed a hypergeometric test to find GO enriched categories. To assign GO terms from the clusters back to the different time points, we require for each time point that at least ¾ of the proteins comprised in each cluster were at least 2-fold enriched relative to their minimum LFQ value. To calculate the time point-specific GO terms, only terms not ubiquitously enriched in all time points were kept. Terms of each cluster were sorted by FDR and similarity scores were calculated. Eventually, only the term with the lowest p-value among similar terms (similarity score higher than 0.7 in a range between 0 and1) was kept.

Life cycle data set: For each stage of the life cycle, we performed a hypergeometric test with the stage-specific proteins.

We additionally performed GO terms analysis on the core proteome (defined as the fraction of proteins detected in all time points) and also scored the similarity between terms. The terms were summarized in a scatter plot and

colored based on their similarity score: 1) we performed hierarchical clustering of the distances between terms, defined as 1-score; 2) we cut the tree into subtrees, grouping together terms with similarity higher than 0.7 (range 0-1, same threshold used before); 3) the dimensionality of the similarity matrix was reduced to 2 dimensions using classical multidimensional scaling, which were used as xy coordinates to distribute the terms in a scatter plot colored based on the cluster assignment; 4) as the cluster representative GO term, the most enriched term (FDR) of each cluster was selected, with the size of the circle representing the number of genes a term contains.

Alternatively, we summarized the GO terms of the core proteome as a Treemap, which is a way of displaying hierarchical data using nested proportional rectangles. In our case, we colored (grouped) the terms based on the subtrees of the previously calculated hierarchical clustering of the similarity matrix. The cluster representative and the size of the rectangle were assigned as described above for the scatter plot.

Integration of RNA-seq data

RNA quantification comes from the Supplementary table S10 of Graveley et al., 2011, titled FPKM levels for FlyBase 5.12 Transcripts from short poly(A)+ RNA-Seq. We used this table to estimate the gene expression naively summing the FPKM values of the different transcripts quantified. In order to control the variance, the FPKM values were log-transformed.

We obtained from Ensembl 84 (Yates et al. 2016) the corresponding FlybaseName Gene ids to the Flybase Transcript ids, which were later used to merge the RNA and protein quantitation.

The plot comparing the similarity between RNA and protein expression at each time point is a false color image matrix with the pairwise Pearson correlation coefficients of all time points between the RNA and protein quantitation.

Translational delay and identification of RNA to protein translation patterns
As part of the integration of RNA and protein data, we grouped together genes that have similar RNA and protein profiles. To do this, we calculated a PCA for the protein and RNA data sets in order to accumulate the maximal variability in one dimension. The first component of the two reduced data sets was then used to cluster the genes by a k-means clustering (max. 1000 iterations, 100 starting random centers). Based on Silhouette plots we chose 6 clusters that explained best our data.

Integration of the Fly-FISH data

We downloaded the localization data for embryos from http://fly fish.ccbr.utoronto.ca/terms/ and complemented them with further annotation from the Ensembl release 84 (ensembl_gene_id, flybasename_gene) using Biomart. Proteins were classified into tissues based on FISH data at not only embryo stages but also taking into account RNA expression in later developmental stages. Genes were automatically clustered based on the LFQ data available, using the Affinity Propagation method described before with negative squared Euclidean distances. The average expression trend of each cluster was then calculated by lowess smoothing (Cleveland 1979, 1981). Eventually, we calculated the GO terms enriched per cluster using the R packages GSEABase, GOstats and org.Dm.eg.db as previously described.

SUPPLEMENTAL TABLES

Supplemental Table 1. Filtered MaxQuant output table with calculated and imputed LFQ values.

Supplemental Table 2. GO enrichment information on the life cycle proteome.

Supplemental Table 3. Calculated dynamicity score table.

Supplemental Table 4. Clusters obtained for the complete life cycle experiment.

Supplemental Table 5. GO enrichment information on the life cycle clusters.

Supplemental Table 6. One week old male and female protein comparison.

Supplemental Table 7. Young male and female protein comparison.

Supplemental Table 8. Overview of maternally loaded RNA and proteins.

Supplemental Table 9. Filtered MaxQuant output table obtained by including ncRNA sequences.

Supplemental Table 10. Filtered MaxQuant output table obtained by including pseudogene sequences.

Supplemental Table 11. Filtered MaxQuant output table with calculated and imputed LFQ values for the embryogenesis time course.

Supplemental Table 12. Automatically generated 70 clusters for the embryogenesis proteome.

Supplemental Table 13. GO enrichment information on the embryogenesis time-course analysis clusters.

Supplemental Table 14. Tissue-specific cluster information.

Supplemental Table 15. GO enrichment information on the tissue-specific clusters.

Supplemental Table 16. Six RNA and protein clusters based on the first PCA component.

SUPPLEMENTAL REFERENCES

Badoahrani S, Hilliker AJ. 2008. Antioxidants cannot suppress the lethal phenotype of a Drosophila melanogaster model of Huntington's disease. Genome 51: 392–395.

Bodenhofer U, Kothmeier A, Hochreiter S. 2011. APCluster: an R package for affinity propagation clustering. Bioinformatics 27: 2463–2464.

Carlson M. 2016. org.Dm.eg.db: Genome wide annotation for Fly.

Cleveland WS. 1979. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association 74: 829–836.

Cleveland WS. 1981. Lowess - A program for smoothing scatterplots by robust locally weighted regression. American Statistician 35: 54.

Cohen J. 1988. Statistical power analysis for the behavioral sciences (2nd ed.) L. Erlbaum Associates, Hillsdale, New Jersey

Damgaard C, Weiner J. 2000. Describing inequality in plant size or fecundity. Ecology 81: 1139–1142.

Falcon S, Gentleman R. 2007. Using GOstats to test gene lists for GO term association. Bioinformatics 23: 257–258.

Frey BJ, Dueck D. 2007. Clustering by passing messages between data points, Science 315: 972–976.

Gini C. 1912. Variabilità e mutabilità, contribrito allo studio delle distribuzioni e delle relazioni statistiche. Fascicolo Ier: Introduzione - Indici di variabilità - Indici di mutabilità. Bologne.

Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG, van Baren MJ, Boley N, Booth BW et al. 2011. The developmental transcriptome of Drosophila melanogaster. Nature 471: 473-479.

Kappei D, Butter F, Benda C, Scheibe M, Draskovic I, Stevense M, Novo CL, Basquin C, Araki M, Araki K, et al. 2013. HOT1 is a mammalian direct telomere repeat-binding protein contributing to telomerase recruitment. EMBO J 32: 1681–1701.

Kondo S, Ueda R. 2013. Highly improved gene targeting by germline-specific Cas9 expression in Drosophila. Genetics 195: 715–721.

Liu N, Lasko P. 2015. Analysis of RNA interference lines identifies new functions of maternally-expressed genes involved in embryonic patterning in Drosophila melanogaster. G3 (Bethesda) 5: 1025–1034.

Morgan M, Falcon S, Gentleman R. 2016. GSEABase: Gene set enrichment data structures and methods.

Schlicker A, Domingues FS, Rahnenfuehrer J, Lengauer T. 2006. A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinformatics 7: 302.

Storey D, Bass AJ, Dabney A, Robinson D. 2015. Q-value estimation for false discovery rate control.

Sury MD, Chen JX, Selbach M. 2010. The SILAC fly allows for accurate protein quantification in vivo. Mol Cell Proteomics 9: 2173-2183.

Tai YC. 2007. Timecourse: Statistical Analysis for Developmental Microarray Time Course Data.

Yates A, Akanni W, Amode MR, Barrell D, Billis K, Carvalho-Silva D, Cummins C, Clapham P, Fitzgerald S, Gil L et al. 2016. Ensembl 2016. Nucleic Acids Res. 44: D710-D716.

Yu G, Li F, Qin Y, Bo X, Wu Y, Wang S. 2010. GOSemSim: an R package for measuring semantic similarity among GO terms and gene products, Bioinformatics 26: 976–978.

How-To for the web interface

This is a short manual on how to navigate the web interface accompanied with the manuscript "The developmental proteome of Drosophila melanogaster".

Contents

1 Data sets

In the top menu [1], you can select the data set to work with: "Whole Life Cycle" or "Embryogenesis".

1.1 Whole life cycle

The data set contains the following 15 time points (measured in quadruplicates):

E02: 0-2 hours old embryos
E06: 4-6 hours old embryos
E12: 10-12 hours old embryos
E20: 18-20 hours old embryos
L1: 40-44 hours after egg laying

L2: 66-68 hours after egg laying
eL3: 83-85 hours after egg laying
L3c: crawling larvae
P1: white pupae
P2-P5: every 24 hours after pupation (white pupae)
Aym/Ayf: adult young male/female flies 4 hours after hatching
Am/Af: 1 week old adult male/female flies

1.2 Embryogenesis

Embryos were collected in quadruplicates with a 30 min laying window for the following time points: 0h, 1h, 2h, 3h, 4h, 5h, 6h, 8h, 10h, 12h, 14h, 16h, 18h, 20h.

2 Web interface structure

2.1 Search

After selecting one of the data sets, a table with all detected proteins will appear. The protein of interest can be queried in the "Search" bar [2] and selected from the table by clicking [3]. After selecting one entry, the protein will be saved in the "update selection" window on the upper right site.

2.2 Selected proteins

This feature allows keeping information about previous searches. For a completely new search, please delete the entry in the upper right window [4] and click the update selection button [5].

2.3 Data analysis

After selecting the protein of interest, the data can be chosen from a second menu [6] with the following tabs:

Proteomics
Proteomics/RNA

Similar protein profiles

Scatter plots

Clusters

All the graphical representations are navigable by hovering over the dots/lines. Additionally, it is possible to download any graphical representation via the "download" button.

2.3.1 Proteomics

Bar plots representing protein abundance (y-axis) throughout the collected time points of the selected data set (x-axis). Colored dots indicate the different replicates. The grey scale of the bars represents the number of replicates in which the protein was measured (grey=1, black=4). The value type legend specifies the origin of the data (for detailed information, please check supplemental methods). In each plot two thresholds are presented: 1% (continuous line) and 99% (dashed line) of the LFQ intensities found in the complete data set. The orange line indicates the tendency of protein expression based on measured and imputed values.

2.3.2 Proteomics/RNA

Protein and RNA profiles of the selected protein are shown. The RNA data is retrieved from the modENCODE project (Graveley et al., 2011). All transcripts for each protein are color-coded and by hovering over the lines, their gene names are displayed.

2.3.3 Similar protein profiles

This feature allows searching for proteins with similar expression profiles to the protein of interest. The 100 most similar proteins are listed in the table on the left (scored by similarity) and can be searched in the search tab. By selecting proteins, expression profiles of protein and RNA will be added to the graphical output in a different color.

2.3.4 Scatter plots

Different scatter plots from the manuscript are available for interactive searches (see left menu for selection). The selected proteins from the search (if available in the scatter plot) are highlighted.

2.3.5 Clusters

Significantly changing whole life cycle protein profiles were clustered and are shown by stage (embryo, larva, pupa, adult). The cluster of interest can be selected and respective protein profiles will be visualized. The clusters that contain information about the selected proteins will be displayed (lower right). Embryogenesis data was complemented with tissue-specific expression data. Therefore, a second tab "tissue" allows the selection of tissue and cluster number showing the respective protein expression profiles.

3 Advanced settings

This tab allows modifying different aspects of the graphical output:

y-axis: the ordinate range can be defined [7]

Value types [8]:

measured: calculated LFQ intensities

missing: no LFQ intensity calculated, but measured intensity

imputed: calculated by imputation (see supplemental methods)

dropped: a single measured replicate value without intensities in neighboring time points was dropped and replaced by an imputed value

Show/hide calculated mean (orange line).

IV. THE RNA FOLD INTERACTOME OF EVOLUTIONARY CONSERVED RNA STRUCTURES IN *S. CEREVISIAE*

The RNA fold interactome of evolutionary conserved RNA structures in *S. cerevisiae*

Nuria Casas-Vila[1,2], Sergi Sayols[1], Lara Pérez Martínez[1], Marion Scheibe[1] and Falk Butter[1,*]

[1] Institute of Molecular Biology (IMB), Mainz, Germany
[2] Current address: ISGlobal, Hospital Clinic –Universitat de Barcelona, Barcelona 08036, Catalonia, Spain.
[*] To whom correspondence should be addressed: f.butter@imb.de; Institute of Molecular Biology (IMB), Ackermannweg 4, 55128 Mainz

ABSTRACT

RNA-binding proteins play key roles in regulation of gene expression via recognition of structural features in RNA molecules. Here we apply a quantitative RNA pull-down approach to 186 evolutionary conserved RNA structures and report 162 interacting proteins. Unlike global RNA interactome capture, we associate individual RNA structures within messenger RNA with their interacting proteins. Of our binders 69% are known RNA-binding proteins, whereas some are previously unrelated to RNA-binding and do not harbor canonical RNA-binding domains. While current knowledge about RNA-binding proteins relates to their functions at 5' or 3'-UTRs, we report a significant number of them binding to RNA folds in the coding regions of mRNAs. Using an *in vivo* reporter screen and pulsed SILAC, we characterize a subset of mRNA-RBP pairs and thus connect structural RNA features to functionality. Ultimately, we here present a generic, scalable approach to interrogate the increasing number of RNA structural motifs.

INTRODUCTION

RNA-binding proteins (RBPs) are key players in several aspects of co- and post-transcriptional gene expression regulation, namely RNA splicing, capping, polyadenylation, export, translation and turnover[1]. While protein-centric methods are widely employed to study RNAs associated with preselected RBPs, RNA-centric methods coupled to mass spectrometry such as RNA pull-downs and RNA interactome capture (RIC) allow identification of protein interactors at a target RNA[2]. Previous polyA RIC studies have catalogued a comprehensive repertoire of RBPs in budding yeast, extending the known set of RBPs from previous low-throughput studies and *in silico* predictions based on similar RNA-binding domains (RBDs)[3,4,5].

RNA can adopt complex structures critical for binding to proteins that can vary at different cellular environments. Thus, intensive efforts have been undertaken to investigate RNA folding. In this context, the *in vivo* identification of RNA structures has been facilitated by techniques such as

selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE) and dimethyl sulfate-sequencing (DMS-Seq)[6]. The latter was also employed to identify structured mRNA regions in budding yeast[7]. Subsequent phylogenetic analysis revealed that 188 of these in vivo structured mRNA regions were conserved among yeast species[7]. However, to further characterize these conserved putative protein-interacting structures a streamlined approach is required that can investigate dozens of short RNA fragments. Employing SILAC-based quantitative mass-spectrometry, we map RBPs for this previous published set of evolutionary conserved RNA structures in yeast, extending the structural information to a functional context and thus providing a generic workflow to systematically study RNA-protein interactions at a large number of RNA folds harboring putative protein recognition elements.

Comparison with previous RBP datasets shows that a set of our RNA fold-associated proteins comprises well-studied RBPs for which we are able to reveal target mRNAs. Furthermore, we also report proteins with unassigned roles in RNA biology. Integration of genetic interaction data and experimental approaches to evaluate translational control provide first hints into the functional consequences of the reported mRNA-RBP interactions.

RESULTS

Identification of proteins enriched at conserved RNA folds

A recent study investigated mRNA structural features using DMS-Seq and revealed hundreds of *in vivo* and *in vitro* structured mRNA regions in *S. cerevisiae*[7]. In order to select for candidate structures with possible biological implications, a cut-off criteria was applied to DMS-Seq signal based on previously known functional structures like HAC1, RPS28B and ASH1. Additionally, phylogenetic analysis on these mRNA structures revealed a list of 188 structured regions under positive evolutionary selection, lending additional support for a physiological function. These 188 evolutionary conserved RNA regions are of similar length, but found in different regions of the transcript (**Supplementary Fig. 1a,b**). We reasoned that one possibility

for their strong evolutionary conservation is the recognition by -proteins (RBPs). When we performed comparative analysis of the provided DMS-Seq datasets, we observed similar *in vivo* and *in vitro* folding features for this set of evolutionary conserved folds, suggesting a very robust intrinsic structural conformation (**Supplementary Fig. 1c**). This high *in vitro*/*in vivo* correlation indicated that our previously developed streamlined RNA-protein interaction screen based on quantitative proteomics[8] is suited to identify putative protein binders to these RNA folds. We were able to transcribe 186 of the 188 conserved RNA folds with a S1 aptamer[9] fused at their 3'-end for immobilization on a streptavidin matrix (**Fig. 1a**). Each of these RNA folds was compared to a generic control bait, a 161-bp 3'UTR *COX17* mRNA fragment harboring two Puf3 binding sites, in a quantitative SILAC-based RNA pull-down. To this end, the fold and the control were incubated with differentially labeled SILAC-encoded extract from *S. cerevisiae*. In order to reduce the number of competing unspecific binders, a pool of all investigated RNA folds without the S1 aptamer was added to each pull-down (**Supplementary Fig. 1d**). In a proof of concept, we applied this workflow to a functionally validated hairpin structure in the 5'UTR of the *PMA1* mRNA[7]. This structure is under positive evolutionary selection as exemplified by compensatory mutations in other yeast species (**Fig. 1b**). Disruption of the stem loop by non-compensatory mutations is known to change gene expression[7]. To explore a possible regulation by RNA-binding proteins, we determined binding partners to this structure and applied our quantitative proteomics workflow comparing the wildtype *PMA1* hairpin to a mutated dysfunctional structure. In this experiment, we identified Sbp1, a known translation repressor, enriched at the wildtype *PMA1* structure, attesting our ability to identify protein-binding partners to RNA structures (**Fig. 1c**). We conducted the screen (744 pull-downs) in a label-switch fashion, resulting in a forward and reverse experiment for each query RNA fold[8]. We compared two strategies to filter for enriched proteins, one with a flexible cut-off depending on the enrichment of the known binder Puf3 on the control bait and the other based on a log2 SILAC

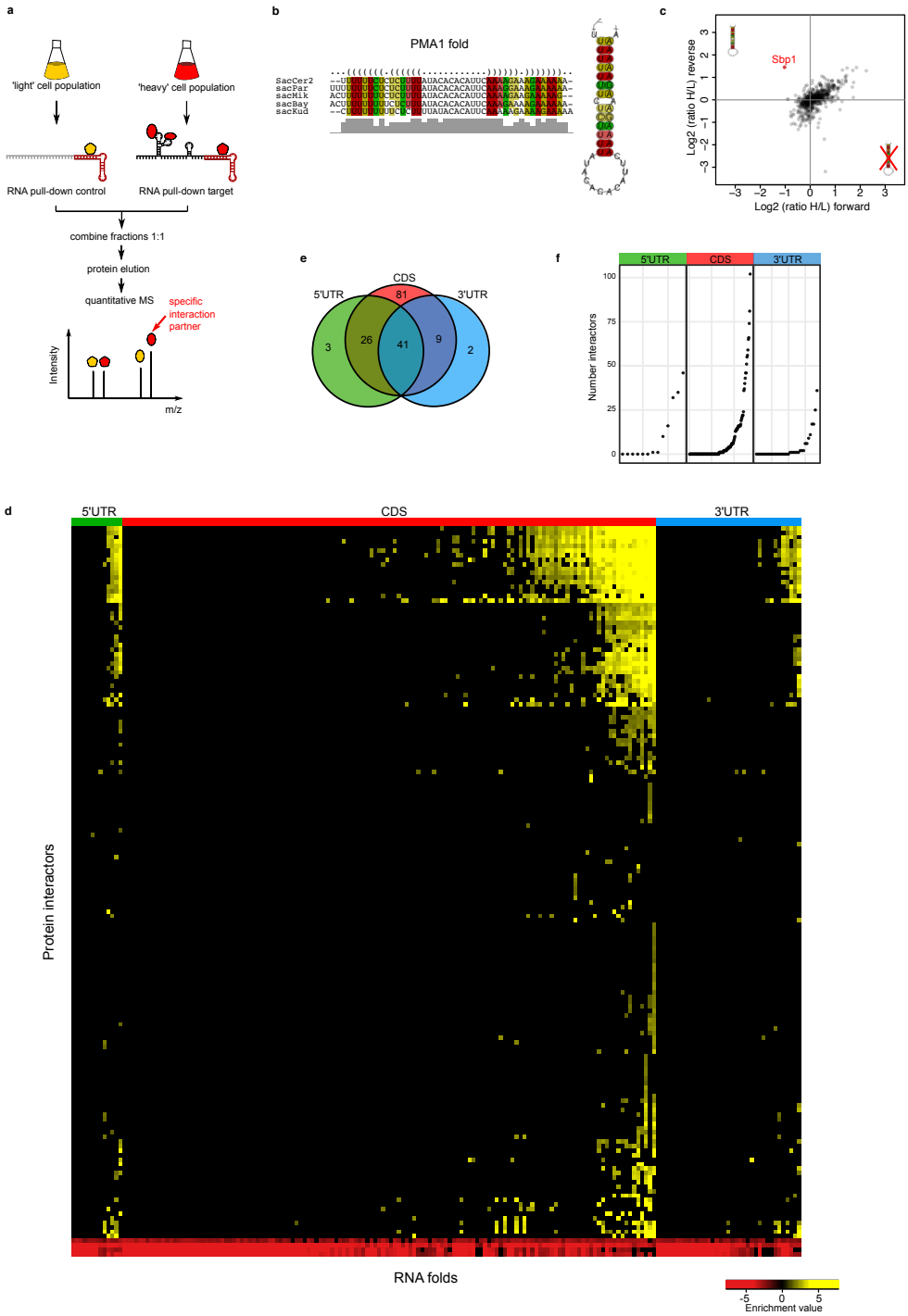

a
'light' cell population
'heavy' cell population
RNA pull-down control
RNA pull-down target
combine fractions 1:1
protein elution
quantitative MS
specific interaction partner
Intensity
m/z
b
PMA1 fold
SacCer2
sacPar
sacMik
sacBay
sacKud
c
Sbp1
Log2 (ratio H/L) reverse
Log2 (ratio H/L) forward
d
5'UTR
CDS
3'UTR
Protein interactors
RNA folds
Enrichment value
e
5'UTR
CDS
3'UTR
3
26
81
41
9
2
f
5'UTR
CDS
3'UTR
Number interactors

Figure 1. Workflow and results for identification of proteins enriched at evolutionary conserved RNA folds. (a) Schematic of the SILAC-based quantitative RNA-protein interactomics workflow. (b) Conservation analysis of the conserved fold in the *PMA1* mRNA by RNAz. Dots indicate unpaired bases and brackets paired bases. Grey bars represent sequence conservation among the indicated 5 yeast species. Folded structure shown on the right, with positional entropy values ranging from red/yellow (lower entropy) to green/blue (higher entropy). (c) Two-dimensional interaction plot comparing the interactors for *PMA1* wildtype hairpin with the mutated fold. (d) Heatmap showing enrichment values for the 162 protein-binding partners to the 186 investigated RNA folds. (e) Venn diagram showing overlap of interactors according to genomic position of the RNA fold (5'UTR, CDS, 3'UTR). (f) Dot-plot displaying the number of binders identified to each investigated RNA fold grouped by localization within the mRNA (13 5'UTR, 136 CDS and 37 3'UTR RNA folds) (Supplementary Data 2).

ratio > 1, representing a two-fold enrichment (**Supplementary Fig. 1e**). In order to ensure technical quality and reproducibility in our streamlined screen, we monitored the enrichment of the known interactor Puf3 at the *COX17* RNA fragment together with three other repeatedly binding proteins (Lsg1, Sui3 and Gcd11) (**Supplementary Fig. 1f**).

Requiring a stringent filter of at least two-fold enrichment against the control RNA in both forward and reverse experiments, we identified 162 proteins interacting with the investigated 186 conserved RNA folds (**Fig. 1d and Supplementary Data 1**). Notably, the length of the RNA fragment did not correlate with the number of bound proteins, excluding a systematic bias as would be apparent for unspecific background (**Supplementary Fig. 1g**). In fact, the number of interacting proteins per mRNA fold fragment is quite diverse (**Supplementary Fig. 1h**). While 25% of our interactors (n = 41) were enriched at folds irrespective of the functional region of the mRNA (5'UTR, CDS, 3'UTR), 50% of them showed positional binding preferences (n = 82) (**Fig. 1e**). Notably, none of our interactors showed exclusive simultaneous binding to the 5' and the 3'UTR, indicating a strong functional separation of the two different UTRs (**Fig. 1e**). Irrespective of their genomic location, for 42% of the RNA folds we did not detect an interaction partner; however, we observe a preference for CDS RNA folds to present a higher number of interactors (**Fig. 1f**). Some RNA folds in our dataset are partially overlapping in sequence within the same mRNA. A comparative analysis on the interactors

of partially overlapping RNA folds can help delimit the relevant sequence for a given protein-RNA interaction. Our interactomics data covers 36 mRNAs with multiple RNA folds (2-6 folds per mRNA) and thus can be used to gain information on the localization specificity of our interactors. A specific analysis on *SSC1* and *YNL190W* mRNAs shows that RNA folds at different positions along the mRNA have a different set of interacting proteins, allowing us to describe the binding position of these interactors (**Supplementary Fig. 1i**).

Correlating our interactor set with RBP features

We first inspected the biochemical properties of our RBP candidate set to exclude putative technical bias in MS measurement. Neither for the measured proteome nor for the RNA fold interactors did we observe a substantial bias for protein size, length and hydrophobicity when compared to all yeast proteins (predicted proteome) (**Supplementary Fig. 2a**). However, for the RNA fold interactors, we noted a significant shift on the isoelectric point distribution, implying that basic proteins are more prevalent in our interactor set compared to the measured proteome, a distinctive feature of known RBPs (**Fig. 2a**)[10]. Consistent with this observation, we found a high enrichment of the basic amino acids lysine, and to some extend also for arginine, in the amino acid composition of our RNA-binding proteins (**Supplementary Fig. 2a**). 25% of our associated RBPs seem to be very specific and recognize a single RNA fold, whereas a significant fraction (24 out of 186) show a promiscuous binding ability, ranging from 20 to 90 target folds (**Supplementary Fig. 2e**). For instance, Yra1 is among the selective RBPs, detected highly enriched at the two RNA folds on *FAS2* and *SSE1* mRNAs. While it is known that Yra1 drives mRNA export and requires the Dbp2 helicase to unwind RNA[11], our data shows co-enrichment of Dbp9, another DEAD-box RNA helicase that has previously been implicated in rRNA processing[12]. Our co-enrichment, together with the recent identification of Dbp9 as an mRNA-binding protein by other studies[3], might point to a potential

coordinated function of Yra1 with Dbp9, in addition to the previously known function with Dbp2.

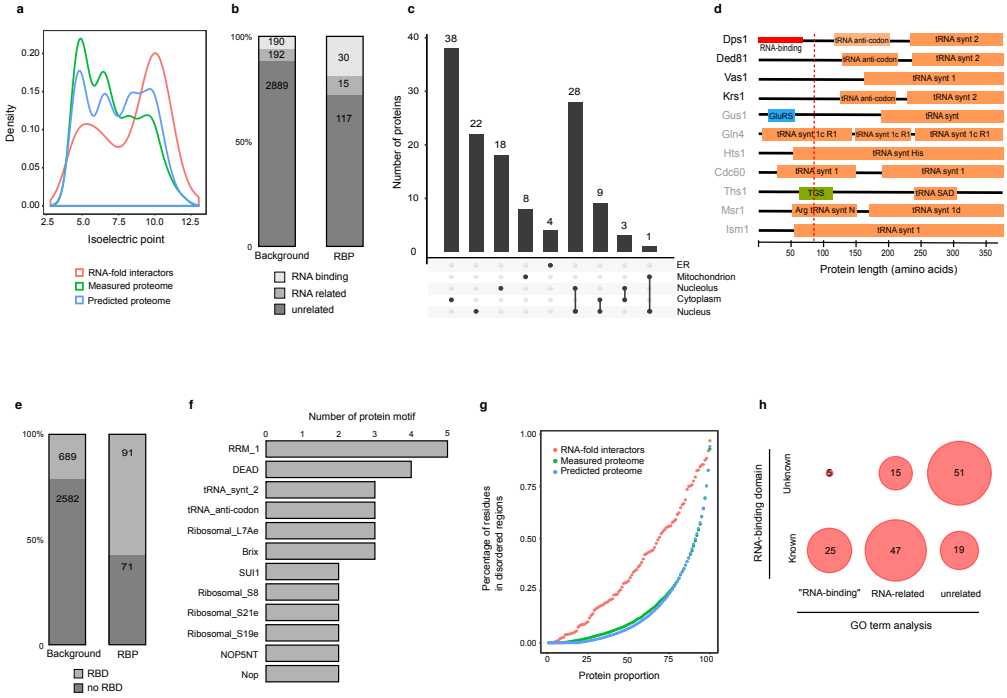


Figure 2. Enrichment of RBP features in our RNA fold interactome. (a) Isoelectric point density of the RNA fold interactome (orange), measured proteome (green) and the predicted proteome (blue) shows bias for RBPs. (b) Classification of RBP candidates based on GO term analysis. Comparison of the measured proteome (background) and the RNA fold interactome. Enrichment of RNA binding and RNA related GO terms was calculated with fisher test odds ratio = 2.9, p-value = $6.2e^{-08}$. (c) Classification of the interactors based on their cellular localization. (d) Schematic of tRNA synthetases domain organization. tRNA synthetases identified in our screen are annotated in black. (e) Count of RBP candidates with RNA related annotated Pfam protein domain (fisher test odds ratio = 4.80, p-value < $2.2e^{-16}$). Shown are the measured proteome (background) and the RNA fold interactome. (f) 12 most abundant RBDs of our RNA fold interactome. (g) Distribution of predicted disordered regions for the RNA fold interactome (orange), measured proteome (green) and predicted proteome (blue) shows enrichment among RBPs. (h) Characterization of RNA fold interactome by GO annotations and protein domain-based classification.

There is no correlation between the protein abundance and the binding of the protein to our RNA folds (**Supplementary Fig. 2b**). We also explored whether high abundant proteins would be biased to bind multiple folds or specific fold types (5'UTR, CDS, 3'UTR) and observed no correlation (**Supplementary Fig.**

2c). GO term enrichment analysis on our interactors ranked by genomic position of the fold (5'UTR, CDS or 3'UTR) revealed enrichment of GTPase activity involved in translation initiation for the 5'UTR interactors, structural constituents of the ribosome and helicase activity for CDS-binding interactors and prevalence of nuclease activity involved in mRNA catabolism for 3'UTR interacting proteins (**Supplementary Fig. 2d and Supplementary Data 6**).

To further characterize our RNA fold interactor dataset, we integrated data from public repositories and previous global RIC screens. As expected, our interactors are enriched for the gene ontology term RNA-binding (**Fig. 2b**). While our approach also identifies proteins that are not in direct contact with RNA, e.g. as part of an RNA-binding protein complex, by using GO annotation, we classify 30 of our 162 interactors as previously reported RNA-binding proteins, such as Bfr1, involved in localization of mRNAs to P bodies[13]. When we classify our interactors according to their cellular localization, we find that 24% are exclusively found in the cytoplasm and approximately 50% relate to nuclear functions (**Fig. 2c**), whereas a few of them localize to mitochondria.

To allow for a more comprehensive examination of our RBP candidates, we extended our computational analysis to include proteins with related GO terms such as tRNA- and rRNA-binding (**Supplementary Data 2**). This increased the number of proteins with a known RNA-related functionality to 45, representing a nearly 3-fold enrichment compared to the reference proteome (**Fig. 2b**).

This includes Dbp3, a RNA-dependent ATPase involved in rRNA cleavage[14] that has previously been reported as an mRNA-binding protein by other RIC studies[3,4,5], four yeast tRNA synthetases Dps1, Ded81, Vas1 and Krs1 and the tRNA ligase Trl1. Examples of tRNA synthetases with a transcript-selective translation control function have been described before in vertebrates and yeast[15,16]. The N-terminus of Dps1 harbors an RNA-binding motif that enables binding to the 5'-end of its own mRNA and thereby inhibits its own translation[15] (**Fig. 2d**). We find that Dps1 binds to 24 of our 186 investigated RNA folds suggesting a broader translational regulation than just its own

mRNA. Indeed, RNA folds bound by Dps1 are primarily located on mRNAs that represent genes with well-defined functions in amino acid synthesis pathways such as MET6, ILV2, ARG1, HIS3 and GUS1[17,18,19,20,21]. Perhaps, in cellular conditions of low amino acid concentrations and impeded tRNA loading, Dps1 becomes available to bind to the conserved RNA folds in the mRNA of these amino acid metabolism genes and thereby controls amino acid synthesis as already suggested for other enigmRBPs[4]. Of note, the three other tRNA synthetases in our interactome (Ded81, Vas1 and Krs1) bind to a small, but highly overlapping set of RNA folds (**Supplementary Fig. 2f**). Interestingly, the target mRNAs of these three proteins are functionally related, as they are involved in protein folding, protein targeting and endocytosis (represented by SSC1, SSE1, VMA3 and SUR7)[22,23,24,25]. This suggests that the possible secondary activity of tRNA synthetases as translation regulators might not be restricted to Dps1. Supporting this idea and in contrast to other yeast tRNA synthetases, the four tRNA synthetases identified in this screen are characterized by a disordered N-terminal extension that might function as an RNA-binding domain[26] (**Fig. 2d**).

We also analyzed the occurrence of classical RNA-binding domains (RBD). To this end, we used Pfam annotations and also evaluated RNA-binding domains from a manual curated dataset[1]. We find that 91 of our interactors have a known RNA-binding domain (**Fig. 2e**). The three previous global RIC studies in *S. cerevisiae* together reported only 48 of these binders while 43 are unique to this study.

Besides the common RBD domains (RRM, DEAD, tRNA- and ribosomal-related) (**Fig. 2f**), we also identify proteins with non-canonical RNA-binding domains. For example, Tma20 binds 11 different RNA folds, has an unknown function but associates with ribosomes, contains a PUA domain and has not been reported by previous interactome capture studies[27]. Proteins with PUA domains might represent a novel type of translation factors, since this domain was detected in proteins that also harbor domains homologous to the translation initiation factors eIF1/SUI1[28,29]. Consistently, Tma20 is

homologous to the human MCT-1 gene that functions as a translation initiation factor[30]. At 8 of our 11 RNA folds targeted by Tma20, we also enrich Tma22, a protein with unknown function, but similar to the human DRP1 protein and possibly linked to translation regulation via its SUI1 domain. Our observations, together with previous studies reporting a physical interaction for these two proteins[7], suggest a possible joint function for Tma20/Tma22 at their target mRNAs. Further experiments are needed to unravel a possible coordinated Tma20/Tma22 function on translation regulation.

Although RBPs have historically been associated with structured RBDs, recent studies revealed protein binding to RNA through intrinsically disordered and low amino acid complexity regions[26]. In this line, a previous interactome capture study from mESCs reported enrichment of low-complexity and disordered regions on their RBP candidates and suggested it to be a general feature of RBPs[31]. Indeed, we also report significant enrichment of disordered region containing proteins on our interactor dataset, further substantiating this as a possible feature of RNA-binding proteins (**Fig. 2g**).

51 of our 162 identified interactors still remain unexplained in the context of RNA associated proteins in *S. cerevisiae* (**Fig. 2h**). Six of these have human orthologues related to RNA biology by GO terms analysis, suggesting a possible RNA function also in yeast. Others, like inosine monophosphate dehydrogenase enzymes (IMD1, IMD2, IMD3) have also been reported in global RIC studies of *S. cerevisiae*[35] and thus may belong to the class of enigmRBPs. While we are not able to clearly identify direct RNA-binding proteins in our setup, most of the additional candidates might also be proteins associating to mRNA in form of complex members, extending the set of proteins involved in RNA regulation beyond the direct interactors reported by RIC (**Supplementary Fig. 2g**).

In vivo validation of mRNA-RBP interactions

We made use of the few available PAR-CLIP data in *S. cerevisiae* to validate our RBP-RNA fold interactions[32]. PAR-CLIP data for Pab1 is consistent with our

results showing Pab1 recognition of RNA folds at the *YEF3* and the *RBG2* mRNA (**Fig. 3a**). We validated this interaction using a TAP-tagged Pab1 strain for the RNA pull-down (**Fig. 3b**). Interestingly, in both cases Pab1 binds to RNA folds in the coding region of its target mRNAs. We used two additional PAR-CLIP datasets for Nab2 and Yra1 to validate even less strong enrichment found for RNA folds within the *RBG2* and *TMC1* mRNA (**Supplementary Fig. 3a).**

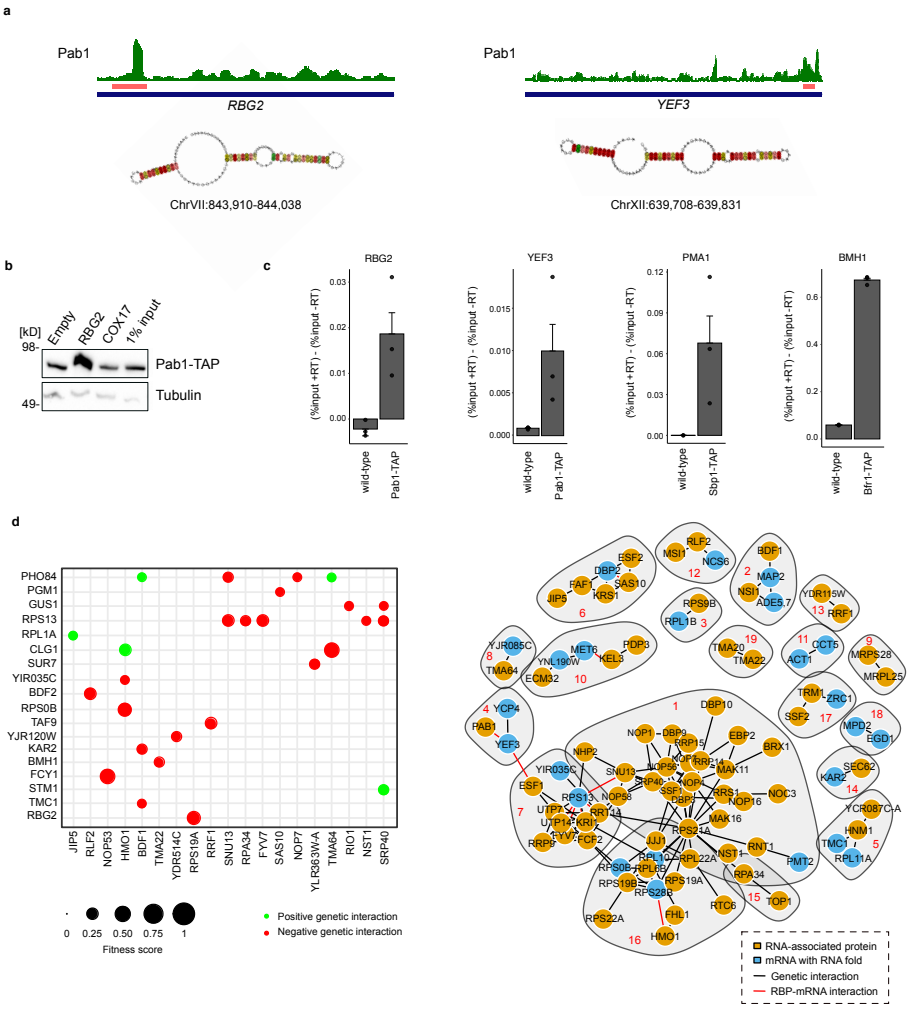


Figure 3. Integration of *in vivo* localization and genome-wide genetic interaction data provides functional insights into our set of interactors. (**a**) Pab1 PAR-CLIP data shows a significant peak on both *RBG2* (p-value = $1.5e^{-14}$) and *YEF3* (p-value = $9.1e^{-9}$) target genes

and overlap with the conserved RNA fold region (colored salmon). Folded structures are shown, with positional entropy values ranging from red/yellow (lower entropy) to green/blue (higher entropy). (**b**) Pull-down and Western blotting on a TAP-tagged Pab1 strain validated Pab1 binding to the *RBG2* RNA fold. (**c**) RNA immunoprecipitation (RIP) experiments in wildtype and endogenously TAP-tagged strains show enrichment of target RNAs in Pab1 (for *RBG2* and *YEF3*), Sbp1 (for *PMA1*) and Bfr1 (for *BMH1*). Results are shown relative to input signals normalized to the –RT (no reverse transcriptase) conditions. Data are presented as mean ±SEM values in n=3 technical replicates. (**d**) Matrix showing genetic interactions described for our RBP (x-axis) and RNA fold (y-axis) interacting pairs. Genetic interactions with a fitness score > 0.08 are colored according to positive (green) and negative (red) interactions. The circle size is proportional to the fitness score of the double knockout strain of the two relevant genes. (**e**) Clustering of our RBP candidate genes and mRNA genes according to the genetic interaction profile correlations (similar genetic interaction profiles considered upon PCC values > 0.2). Red numbers indicate the ID of the respective gene community (**Supplementary Data 4**).

Extending our validation, we performed RNA immunoprecipitation to validate a few more protein-RNA interactions using available TAP-tagged proteins: Pab1 to YEF3 and RBG2, Sbp1 to PMA1 and Bfr1 to BMH1 (**Fig. 3c**).
Overall, these selected examples underscore nicely that we indeed identified in vivo relevant RBP-mRNA interactions.

Functional hints from genetic interaction data integration

In order to explore functional relationships between our RBP candidates and their target mRNAs, we made use of recent whole genome yeast genetic interaction data[33]. Whereas the genetic interactions suggest synergistic effects of genes working in compensatory pathways, combination with our interactomics dataset can point to a possible mechanistic model. Our analysis resulted in 27 positive or negative genetic interactions of our RBP-mRNA pairs that allow speculation of mechanistic links (**Fig. 3d and Supplementary Fig. 3b**). For example, *TMA64* presents a positive genetic interaction with *PHO84* and additionally, we detected Tma64 enriched at the evolutionary conserved RNA fold on *PHO84* mRNA that encodes for an inorganic phosphate transporter (**Supplementary Fig. 3c**). Tma64 is not described at a functional level; however, it harbors a putative RNA-binding domain and has been linked to translation control[34,35]. In addition, *TMA64* presents another genetic interaction with the *CLG1* gene and we report it as interactor of an RNA fold

on the *CLG1* mRNA. Clg1 is a cyclin-like protein that exerts its function through the interaction with Pho85, a cyclin-dependent kinase linked to phosphate response[36]. Specifically, under high phosphate conditions, the Pho85-Pho80 complex phosphorylates the transcription factor Pho4, promoting its nuclear export and thereby preventing transcription of genes related to phosphate starvation[37] (**Supplementary Fig. 3c**). In addition, we report Tma64 enrichment at a 5'UTR RNA fold of the *PCL5* mRNA, encoding for yet another cyclin that is phosphorylated by Pho85 and involved in the amino acid starvation response[38] (**Supplementary Fig. 3c**). Altogether, collective evidences from our interactomics screen and the genome-wide genetic interaction data insinuate a role for Tma64 in phosphate homeostasis, perhaps regulating the expression of starvation-related genes via recognition of structural features on its target mRNAs.

To further examine functional connections between our RBP and mRNA folds, we clustered them based on similarity in their large-scale genetic interaction profile and tested for enriched GO annotations in each group (**Fig. 3e and Supplementary Fig. 3d**). Genes that belong to the same pathway often share phenotypes with their target genes and therefore share similar genetic interaction profiles. For example, the central densely connected nodes mainly consisting of RBP candidates are characterized by ribosomal-related functions such as rRNA processing and ribosome biogenesis (communities 1, 7, 16). Also, other clusters present mitochondrial translation (community 9) or chromatin organization (community 2) functions. Similarly, the membership of uncharacterized RBP candidates to the resulting communities can be used to infer putative functions. For instance, in addition to the *in vivo* validated binding of Pab1 to the conserved RNA fold on the *YEF3* mRNA (**Fig. 3a**); a functional connection between *PAB1* and *YEF3* mRNA is further supported by similar genetic interaction profiles (community 4) (**Fig. 3e**). In another case, *YDR115W* clusters together with the mitochondrial ribosome-recycling factor *RRF1*, perhaps suggesting a possible role for YDR115W in mitochondrial translation (community 13). As shown above, our integration with physical

interactomics data can provide hints towards a putative regulation mechanism of genes that share similar genetic interaction profiles.

Translational control by RBPs binding to mRNA folds

To explore the biological consequences of our identified RBP-RNA fold pairs experimentally, we first focused on possible translational regulatory effects exerted by RNA folds in UTR regions. We incorporated two 5'UTR- and ten 3'UTR-folds at the respective untranslated regions of a GFP reporter construct and quantified changes in GFP expression by fluorescence intensities upon knockout of the respective interacting RBPs (**Fig. 4a and Supplementary Fig 4a**). We validated our strategy with a known functional 5'UTR fold *SFT2*[7]. Consistent with previous data, disruption of base-pairing interactions in the 5'-*SFT2* hairpin resulted in an increase of GFP levels, demonstrating a repressive function of the *SFT2* hairpin on mRNA expression (**Supplementary**

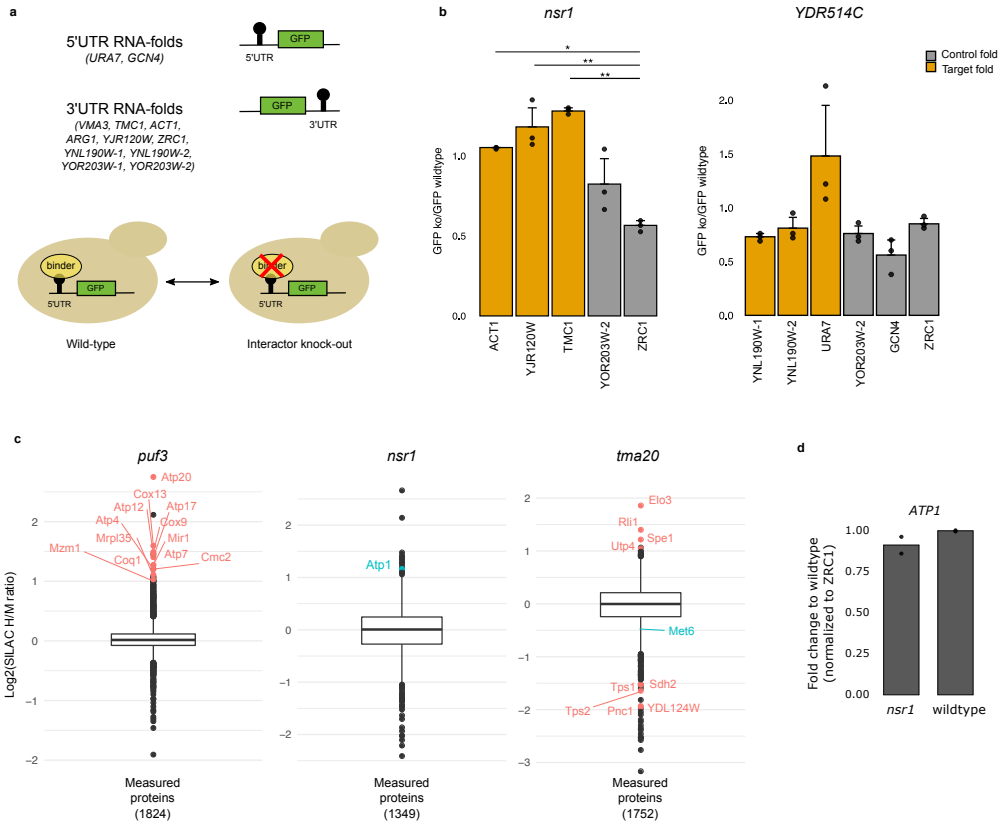


Figure 4. Functional validation of mRNA-RBP interactions. (**a**) Schematics of the GFP reporter screen with either a 5'UTR or 3'UTR fused RNA fold. Expression is compared between the wildtype and knockout yeast strain of the respective interactor. (**b**) Bar-plots display changes on the reporter expression levels when Nsr1 and YDR514C are bound to their respective folds (orange). Control experiments are RNA folds with no detected Nsr1 or YDR514C binding (grey). P-value * < 0.05 and ** < 0.01 based on analysis of variance (two-sided ANOVA and Tukey post-hoc test) in n=3 biologically independent samples. ACT p-value = 0.017, YJR120W p-value = 0.005 and TMC1 p-value = 0.003. Data are presented in mean ±SD values. (**c**) Pulsed SILAC box plots showing log2(SILAC H/M ratio) values of all measured proteins for *puf3* knockout, *nsr1* knockout and *tma20* knockout strains compared to a wildtype strain (**Supplementary Data 7**). Boxes show median (center) and interquartile ranges (ends), lower whisker representing the smallest observation greater than or equal to 1.5 times the interquartile range and upper whisker representing the largest observation less than or equal to 1.5 times the interquartile range. (**d**) qRT-PCR expression levels of *ATP1* mRNA in a *nsr1* knockout strain compared to the wildtype strain. Data are presented as mean values in n=2 technical replicates.

Fig. 4b). We applied this workflow to our 12 different UTR folds, which were investigated in the relevant yeast knockout strains and compared to wildtype (**Fig. 4a**). We quantified reporter levels of all 73 possible combinations of these 12 RNA folds and their identified interactors under normal growth conditions. Two RNA folds without any identified interactor (**Fig. 1c**) were used as controls. While transcriptional regulation is only one possible regulatory scenario that can be executed by our RNA fold interactors, we found two of them Nsr1 and YDR514C, that showed expression changes of the reporter construct (**Fig. 4b**). The yet uncharacterized protein YDR514C that bound to the RNA fold in the 5'UTR of the *URA7* mRNA behaves like a translational repressor in our screen, identical to Nsr1. While YDR514C has no annotated domains, Nsr1 harbors two RNA recognition motifs (RRM) and has been reported to be involved in rRNA processing[39]. In our screen, we detected Nsr1 binding to 60 of our RNA folds (**Fig. 1d**) and tested three of these RNA folds found in the *ACT1*, *YJR120W* and *TMC1* mRNAs, for a putative translational regulation. Concomitant with a role in gene expression, we observed an increase in reporter levels for the three Nsr1 target 3'UTR RNA folds upon knockout of Nsr1, while at two control RNA folds that did not interact with Nsr1 we did not observe increased GFP expression.

We also explored a putative role of selected protein-RNA fold pairs on the translation of their target mRNAs by pulsed SILAC[40]. This technique determines protein synthesis and turnover rates via the differential incorporation of isotopically labeled amino acids. We thus compared global protein translation rates between wildtype and selected strains where our interactor was deleted. We used Puf3, a protein known to promote mRNA degradation of nuclear-encoded mitochondrial proteins[41] as a positive control and indeed observed upregulation of the reported Puf3 target proteins in the Puf3 knockout strains (**Supplementary Fig. 4c**). We additionally detected upregulation of 12 additional mitochondrial proteins (**Fig. 4c**). In 6 cases their mRNAs contain a single or multiple copies of the canonical Puf3 binding motif (UGUAAAUA)[42] in their 5'UTR, CDS or 3'UTR (**Table 1**).

In line with our GFP reporter screen data, pulsed SILAC experiments showed a translational repression role of Nsr1 on its target *ATP1* (**Fig. 4c**), as no changes were observed at the mRNA level (**Fig. 4d**). In agreement with our observations, MS studies[43,35] captured Nsr1 as a physical interactor of proteins with a defined role in post-transcriptional gene regulation such as the RNA helicase Dbp2[11], the translation initiation factor Ded1[44,45] and the RNA poly(A) tail binding protein Pab1[46].

We also investigated Tma20 by pulsed SILAC, the protein we found in our screen to be binding to several RNA folds and that is homologous to the human MCT-1 translational regulator. We found proteins involved in different cellular metabolic processes related to carbohydrates (Tps1, Tps2), nucleotides (Pnc1), amides (YDL124W), Acetyl-CoA (Sdh2) and amino acids (Met6) to be upregulated in its knockout strain compared to wildtype. These data establish Tma20 is a positive translational regulator as suggested before based on our interactome data (**Fig. 1d**).

DISCUSSION

In this study, we have used a streamlined RNA interactomics strategy to map protein interactors to 186 evolutionary conserved RNA folds in *S. cerevisiae*.

We showed that our interactor set was enriched for RBP features and fused our interactomics data with genome-wide genetic interaction data to suggest putative functional and mechanistic insights for the detected mRNA-RBP pairs. Finally, we explored translational regulation as a possible functionality of our mRNA-RBP interactions using a reporter screen system and pulsed SILAC.

While current approaches to study RNA-protein interactions such as RIC globally address RBPs binding to mRNA and polyadenylated lncRNA, we here map 162 RBP interactors to individual RNA folds. We show that our setup captures interactions at these RNA structures even for low abundant RBPs as judged by their low abundance in our measured proteome and that were missed by RIC experiments. Our approach does not only capture proteins in direct contact with RNAs but with the current washing conditions also enriches for RBP complexes. To quantify this ratio based on our analysis: 69% are known direct RBPs, while the remaining might be associated via protein-protein interactions.

We used available PAR-CLIP *S. cerevisiae* datasets to validate our detected RNA fold-RBP interactions. While the number of PAR-CLIP datasets in yeast is extremely limited, the available studies show a very high overlap to our data attesting our ability to report physiologically relevant interactions. We furthermore validated experimentally more mRNA-RBP interactions by RNA immunoprecipitation ourselves.

Current knowledge about RBPs involved in post-transcriptional regulation focuses on either the 5' or 3'UTR sequences. Interestingly, a significant portion of the conserved RNA folds is found in the CDS of mRNAs. We here report a significant number of RBPs recognizing RNA folds in the CDS of mRNAs. These interactors are related to ribosomal biosynthesis, tRNA-binding or are metabolic enzymes and kinases. In addition, the binding profile of some classical RNA-binding proteins can be surprising, such as the binding of Pab1 to CDS RNA folds such as *YEF3* and *RBG2* (cross-validated by PAR-CLIP data). These observations suggest that RNA-binding motifs may also form in the

coding region, possibly allowed by the degenerated codons. Whether this binding results in functional consequences will need to be explored in the future.

As noted in previous RIC experiments already, not all discovered RBPs harbor canonical RBDs, but also include known proteins with diverse cellular functions like kinases, metabolic enzymes and tRNA- and rRNA-metabolism factors[10]. Using the resolution for individual RNA folds, we here showed that tRNA synthetases bind to a specific set of RNA folds. This observation immediately suggests that some tRNA synthetases harbor yet unknown RNA-binding domains and hint to a possible role as translation regulators.

Overall, we here outlined how smaller structural features within an RNA molecule can be investigated with a streamlined assay, resulting in the identification of 162 interactors to 186 evaluated RNA folds. Mutational studies should be the next step to characterize our set of protein-RNA fold interactions to decipher the required RNA-binding motifs in more detail. There will be a growing demand for such analyses as thousands of structures can nowadays be either obtained using *in vivo* RNA structure methods, like DMS-Seq or SHAPE, functional reporter assays or computationally predicted based on conformation energy[47].

METHODS

DMS-Seq data comparison for conserved RNA folds

DMS-Seq datasets from 3 different experimental conditions (*in vivo*, *in vitro* and denatured) were retrieved from GSE45803[7]. DMS signals of the corresponding genomic loci for each RNA fold were retrieved and normalized proportionally to the most reactive base within a given structure. For each RNA fold, denatured DMS signal was subtracted to the *in vivo* and *in vitro* DMS signals and Spearman correlations between the resulting *in vivo* and in *vitro* DMS signals were calculated as a measure of similarity. The probably of each RNA fold to form structure was calculated based on the Gini coefficient. This coefficient defines inequality within a population, which in our case means

that RNA folds with high probability for structure formation will have very unequal DMS signal distribution along the sequence (Gini coefficient ≅ 1), while folds with low structural formation capacity will have an even DMS signal distribution (Gini coefficient ≅ 0). The resulting Gini coefficients of the conserved RNA folds (n = 188) were compared to the Gini coefficients of 200 randomly picked regions of similar average length from different genomic locations (intergenic, 5'UTR, 3'UTR, CDS).

RNA conservation analysis

PMA1 RNA fold sequences for *S. cerevisiae*, *S. paradoxus*, *S. mikatae*, *S. kudriavzevii*, *S. bayanus* were obtained from UCSC (SacCer_Apr2011/SacCer3 assembly) and aligned with ClustalW version 1.81[48]. The resulting multiple sequence alignment was fed into RNAalifold 2.2.8 from the Vienna RNA package[49] to predict the consensus secondary structure.

Yeast genomic DNA extraction

Yeast genomic DNA from BY4741 strain was extracted from a 2 ml saturated culture, centrifuged (2,000 g, 5 min) and resuspended in Lyticase buffer (1 M Sorbitol, 100 mM EDTA pH 8, 14.3 mM beta-mercaptoethanol). After 1 h incubation at 37 °C with 2.5 µL lyticase (Zymolyase® 20T (≥ 20 U/mg)), samples were centrifuged (5,000 rpm, 5 min) and pellets washed twice with 1 ml spheroblast wash buffer (1 M Sorbitol, 100 mM EDTA pH 8). Pellets were then resuspended in 500 µL TE 50/100 (50 mM Tris-HCl pH 7.5, 100 mM EDTA pH 8) and incubated with 50 µL 10% SDS for 30 min at 70 °C. 250 µL of 5 M potassium acetate was added and mixed by pipetting, followed by 15 min incubation on ice. Samples were centrifuged at 14,000 rpm for 20 min and supernatants transferred to a clean tube before isopropanol precipitation (700 µL isopropanol followed by 10 min centrifugation at 14,000 rpm). Pellets were cleaned with 70% ethanol before adding 500 µL TE 10/1 solution (10 mM Tris-HCl pH 7.5, 1 mM EDTA pH 8) followed by 30 min heating at 42 °C and gentle shaking. When DNA was completely dissolved, 50 µL of 3 M sodium

acetate were added, followed by a second isopropanol precipitation round (500 µL isopropanol, centrifugation at 14,000 rpm for 15 min). Samples were cleaned in 70% ethanol and pellets dissolved in 50 µL TE 10/1 (10 mM Tris-HCl pH 7.5, 1 mM EDTA pH 8) during incubation at 42 °C for 1 h. DNA concentration was assessed by A280 absorbance on a Nanodrop instrument (Peqlab).

Engineering primer sequences and cloning

The chromosome coordinates of all conserved RNA folds were retrieved from https://weissmanlab.ucsf.edu/yeaststructures/ and complete sequences extracted using R[50] and the Bioconductor[51] packages "BSgenome.Scerevisiae.USCS.sacCer2" and "BSgenome", which contain the representation of the full genome sacCer2. The resulting multifasta file containing all DNA sequences was parsed and used for forward and reverse amplification primer design taking the first 'n' nucleotides from the 5'- or 3'-end until the melting temperature exceeded 58 °C, calculated as 4*(nC + nG) + 2*(nA + nT). Extracted genomic DNA together with the respective primer pairs were used to amplify each RNA fold in a PCR reaction using OneTaq polymerase according to manufacturer's protocol (New England Biolabs). Successful amplification products were monitored on an agarose gel and were subsequently TA cloned into the pcDNA 3.3 TOPO vector following the manufacturer's protocol (Invitrogen). Alternatively, for RNA folds failing PCR amplification, a primer extension approach using long primer dimers was employed, followed by TA cloning into pcDNA 3.3 TOPO vector. Correct sequence of all amplicons or chemically synthetized baits was checked by Sanger sequencing. For the GFP reporter screen, a yeast centromeric plasmid backbone was modified to incorporate the fluorescent yeast enhanced GFP (yeGFP) reporter gene driven by the Adh1 promoter and terminator. RNA folds were amplified from the pcDNA 3.3 TOPO plasmid with generic primers that introduced the corresponding restriction enzyme cutting sites for subsequent cloning into the GFP plasmid. According to the RNA fold UTR

location, RNA folds were cloned 5' or 3' of the GFP reporter with a short linker region; plasmids were transformed into BY4741 yeast strain and selected by the URA3 marker. For the *SFT2* 3'UTR experiments, only the hairpin sequence as previously described[7] was inserted at the 3'-end of the GFP reporter.

SILAC labeling of *S. cerevisiae* and extract preparation

Yeast strain YAL6B auxotrophic for arginine and lysine was grown at 30°C to stationary phase in YPD media and then inoculated 1:10,000 into self-made filter-sterilized SILAC media (6.7 g L^{-1} YNB without aminoacids, 2 % Dextrose, 200 mg L^{-1} L-adenine sulphate, 100 mg L^{-1} L-tyrosine, 10 mg L^{-1} L-histidine, 60 mg L^{-1} L-leucine, 10 mg L^{-1} L-methionine, 60 mg L^{-1} L-phenylalanine, 40 mg L^{-1} L-tryptophane, 20 mg L^{-1} Uracil, 20 mg L^{-1} L-arginine (all amino acids from Sigma-Aldrich) and 30 mg L^{-1} of either 'light' [$^{12}C_6$,] or 'heavy' [$^{13}C_6$]L-lysine (Euriso-top). Cells were grown for at least 10 doublings to allow complete labeling and harvested at exponential growth by centrifugation at 20,000 rpm for 45 min at 4 °C (SORVALL ultracentrifuge, Thermo). 2-liter yeast pellets were resuspended in 30 mL lysis buffer (50 mM Tris-HCl pH 7.5, 100 mM NaCl, 1 mM EDTA, 1 mM DTT, 5% glycerol, 1 mM PMSF, 1 µg mL^{-1} Leupeptin, 1 µg mL^{-1} Pepstatin A) and lysed three times at 35,000 psi in a French press at 4 °C. Different batches of lysate preparation were pooled for homogenization and successful incorporation of labeled lysine was checked by mass spectrometry. Extracts for Western blotting analysis were prepared in lysis buffer (100 mM NaCl, 50 mM Tris-HCl pH 7.5, 10 mM $MgCl_2$, 1 mM PMSF, 0.01% IGEPAL CA-630) by bead milling using 0.5-mm Zirconia/Silica beads in a FastPrep for three cycles of 30 sec at 4 °C with 1 min rest in between. Extract protein concentration was determined by Bradford (BioRad).

RNA transcription and RNA pull-down

The RNA fold baits were created by a PCR reaction with generic amplification primers using the pcDNA 3.3 TOPO plasmid as a template. The forward primer (5'- CGTTAATACGACTCACTATAGGGATCGAACCCTT-3') incorporated the T7

promoter sequence at the 5'-end of the RNA fold amplicon and the reverse primer (5'-CATGGCCCGGCCCGCGACTATCTTACGCACTTGCATGATTCTGGTCGGTCCCATGG ATCCAAAAAAAGATCGAACCCTT-3') added the S1 minimal aptamer sequence at the 3'-end[9]. PCR products were used for *in vitro* transcription was performed according to the manufacturer's protocol (Fermentas) and successful transcription monitored by agarose gel electrophoresis. Tagged RNA oligonucleotides were purified with G-50 micro spin columns (GE Healthcare) and concentration assessed by A280 absorbance on a Nanodrop system (Peqlab). 25 µg of each S1-tagged RNA fold was coupled to paramagnetic streptavidin C1 beads (Dynabeads MyOne, Invitrogen) in RNA binding buffer (100 mM NaCl, 10 mM $MgCl_2$, 50 mM HEPES-KOH pH 7.4, 0.5 % IGEPAL CA-630) and incubated on a rotating wheel for 30 min at 4°C. RNA-bound beads were washed 3 times with RNA washing buffer (250 mM NaCl, 10 mM $MgCl_2$, 50 mM HEPES-KOH pH 7.4, 0.5 % IGEPAL CA-630), followed by incubation with 400 µg yeast extract for 30 min at 4°C on a rotating wheel. At this point, 650 ng of a competitor mixture containing a pool of all RNA baits without the S1-aptamer tag was added to reduce the number of sticky protein binders. After mild washing, light and heavy fractions were combined and samples were boiled in 1x LDS buffer (Invitrogen) and separated on a 4-12% NuPAGE Novex Bis-Tris precast gel (Life Technologies) at 180V in 1x MOPS buffer.

Western blotting analysis

20 µg of whole lysate were run on a 4-12% NuPAGE Novex Bis-Tris precast gel (Life Technologies), transferred to a Protran 85 membrane (Whatman) and probed with either a rabbit TAP antibody (Thermo, 1:1,000) or mouse GFP (Roche, 1:1,000) as primary antibodies. Either rabbit or mouse HRP-conjugated antibodies (GE Healthcare, 1:2,000) were used for detection. Chemiluminiscence detection was done using a SuperSignal West Pico

solution (Pierce) and the SeeBlue Plus2 Pre-stained Protein Standard (Thermo Scientific) was used as a marker.

MS sample preparation and measurement

Coomassie stained gels were cut in one slice and destained with 50% EtOH/25 mM ammonium bicarbonate (ABC). The resulting gel pieces were dehydrated with 100% acetonitrile (ACN) and dried for 5 min in a concentrator (Eppendorf). Samples were incubated with reduction buffer (10 mM DTT/50 mM ABC) for 30 min at 56 °C and further alkylated for 30 min in the dark with iodoacetamide (50 mM IAA/50 mM ABC). Gel pieces were completely dehydrated with ACN and covered in LysC solution (1 µg LysC per sample). Proteins were digested overnight at 37 °C and peptides were extracted twice by incubation with extraction buffer (3% TFA and 30% ACN) for 15 min. The gel pieces were dehydrated with 100% ACN and the extracted volume reduced to approximately 150 µl in a concentrator (Eppendorf). Extracted peptides were desalted in StageTips[52] using two layers of C_{18} material (Empore). Eluted peptides were injected via an autosampler into an uHPLC (EASY-nLC 1000, Thermo) and loaded on a 25 cm capillary (75 µm inner diameter; New Objective) packed in-house with Reprosil C18-AQ 1.9 µm resin (Dr. Maisch) for reverse-phase chromatography. The EASY-nLC 1000 HPLC system was directly mounted to a Q Exactive Plus mass spectrometer (Thermo). Peptides were eluted from the column with a 90 min optimized gradient from 2 to 40% ACN with 0.1% formic acid at a flow rate of 200 nL min^{-1}. Chromatography was stabilized with a column oven set-up operating at 40 °C (Sonation). The heated capillary temperature was set to 250 °C. Spray voltage ranged from 2.2–2.4 kV. The mass spectrometer was operated in data-dependent acquisition mode with one MS full scan and up to ten triggered MS/MS scans using HCD fragmentation[53]. MS full scans were obtained in the orbitrap at 70,000 resolution with a maximal injection time of 20 ms, while MS/MS scan resolution was set to 17,500 resolution and maximal injection for

120 ms. Unassigned and charge state 1 were excluded from MS/MS selection and peptide match was preferred.

MS data analysis

Raw files were processed with MaxQuant (version 1.5.2.8.)[54] and searched against *Saccharomyces cerevisiae* Ensembl annotated protein database R64-1-1.24 Oct 2014 (6,692 entries) using the Andromeda search engine[55]. Carbamidomethylation was set as a fixed modification, while acetyl (N-term protein) and oxidation (Met) were considered as variable modifications. LysC (specific) was selected as enzyme specificity with maximal two miscleavages for MaxQuant analysis. Proteins were quantified with at least 2 ratio counts based on unmodified unique and razor peptides. Known contaminants and reverse hits were removed before plotting the protein ratios of the forward and reverse experiments in R (version 3.2.2). Protein interactors for each RNA fold were identified requiring an enrichment of two-fold in both forward and reverse experiment (log2 SILAC ratios > 1). The enrichment value for each mRNA-protein pair was calculated as the log2 of the Euclidean distance to the origin (coordinates 0,0 in an Euclidean space build upon the dimensions defined by the forward and reverse experiments for each RNA fold). Alternatively, another selection method is proposed to identify interaction partners (Supplementary Figure 1e): proteins showing enrichment higher than the distance to the origin of a known positive control Puf3.

GO annotations and RNA-binding domain analysis

Our interactors were catalogued as RNA-binding when their associated GO term for Molecular Function (Ensembl version 86, Oct 2016) contained the string RNA binding. For those interactors that did not relate to the term RNA binding, a second classification was done as RNA related based on previously described RNA related GO term annotations[1]. For yet unclassified interactors, their human homologs were classified with the same RNA binding and RNA related annotation criteria. As a control, proteins from a whole cell lysate

measurement were equally classified. For RNA-binding domain classification, a curated list of known PFAM RNA-binding domains[1] was used (Supplementary Data 3).

GO annotation of interactors

We performed GO enrichment analysis of binders using SGD Slim Mapper tool. Binders were classified in 3 groups according to the genomic position of the RNA fold they bind to (5'UTR, CDS, 3'UTR).

Biochemical properties analysis of our RBP set

A peptide properties table containing information about Molecular Weight, Isoelectric Point, Protein Length, Hydropathicity GRAVY Scores, Aromaticity Score (frequency of aromatic amino acids: Phe, Tyr, Trp), Codon Adaptation Index, Codon Bias, FOP Score (Frequency of Optimal Codons), Instability Index and Aliphatic Index, was downloaded from the Saccharomyces Genome Database[56]. Significant differences of our RBP set against two control groups containing all known proteins of the yeast genome (predicted proteome) and our measured proteome were calculated with a Student's *t*-test (p-value corrected for multiple testing, FDR).

Disordered region analysis

Complete peptide sequences for our interactor set were retrieved from Ensembl version 86 (Oct 2016) and used for disordered region probability calculation with IUPred2A[57], defined as a lack of known tertiary structure under native conditions. The default prediction type for long disorder region was used and a score based on the percentage of bases with disorder probability higher than 50% was calculated for each protein interactor.

Genetic interaction data integration and network analysis

For genetic interaction data integration, a recent large-scale genetic interaction study[33] was used. For genetic interaction scores (|E|) higher than 0.08 between our RNA folds and protein interactors, an annotation matrix was

calculated depicting the reported |E| scores. Pearson correlation coefficients (PCC) of genetic interaction profiles for all relevant genes (including protein interactors and their mRNA targets) were calculated. Genes with similar genetic interaction landscape were identified using a simple network analysis: each gene was a vertex, and an edge between 2 vertices was defined if the PCC between these two proteins was higher than 0.2, as described in Costanzo; M, et al. Community structure on the network was inferred using a method implemented in the igraph package[58] based on propagating labels, which works by assigning vertices to unique communities and then updates those communities by doing majority voting around a vertex.

PAR-CLIP data validation

PAR-CLIP raw data for Nab2 (GSM1442550), Pab1 (GSM1442553) and Yra1 (GSM1442559) were downloaded from GSE59676[59]. Reads were preprocessed with the Fastx toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) to remove adapter sequences, keep sequences longer than 15 nucleotides, filter artifacts and remove low quality reads (Phread scores < 23). Preprocessed reads were then aligned onto the sacCer3 reference genome with Bowtie[60] version 1.1.2 and options -q -p 8 -S -v 2 -m 10 --best --strata to allow up to 2 mismatches and reads mapping to up to 10 loci, keeping only alignments in the best stratum. Scaled BigWig tracks were generated from the alignment files. To call peaks, the Piranha[61] 1.2.1 peak caller was used with options -b 50 -d Poisson, which bins reads into bins of 50 bp and uses the Poisson distribution to model the counts. Peaks called with Benjamin and Hochberg (BH) corrected p-value below 0.05 were considered true peaks.

Yeast tagged- and knockout strains

Yeast knockout and TAP-tagged strains used for pSILAC, WB and RIP experiments are listed in Supplementary Data 5. All strains were validated by PCR prior to experiments.

RNA immuniprecipitation (RIP) analysis

TAP-tagged strains from Dharmacon collection were used for TAP-RIP experiments. 100-150 ml of exponentially growing cultures were cross-linked for 10 min with 1.2% formaldehyde (Applichem) after cell number normalization. Samples were quenched with glycine (360 mM, Applichem) for 5 min at room temperature. After cooling down to 4 °C on ice for 15 min, cells were pelleted at 4 °C by centrifugation (1731 rcf, 3 min), washed twice with ice-cold PBS and stored at -80 °C until processing. Cell pellets were lysed in FA buffer (50 mM HEPES-KOH pH 7.5, 140 mM NaCl, 1 mM EDTA pH 8, 1% Triton X-100, protease inhibitor cocktail (Roche) via two 30 s rounds of 6.5 M/s FastPrep (MP Biomedical). Samples were diluted in FA buffer supplemented with 0.1% sodium-deoxycholate (SOD). Soluble and chromatin extracts were separated by centrifugation (7 min at 17949 rcf). Subsequently, 2 mg of soluble extracts were incubated overnight at 4 °C with 75 µl of pre-washed IgG Beads (GE Healthcare) with 5% BSA. 50 µl of extracts was separated as an input control. Beads were washed with 1ml of FA buffer, Buffer 500 (50 mM HEPES-KOH pH 7.5, 500 mM NaCl, 1 mM EDTA pH8, 1% Triton X-100, 0.1% SOD), Buffer III (10 mM Tris-HCl pH 8, 1 mM EDTA pH 8, 150 mM LiCl, 1% NP40, 1% SOD) and TE buffer (100 mM Tris-HCl pH 8, 50 mM EDTA pH 8) at 4 °C with 5 min incubation times between washes. Proteins were eluted with Elution Buffer (50 mM Tris-HCl pH 7.5, 1% SDS, 10 mM EDTA pH 8) twice for 8 min at 65 °C. Samples were de-crosslinked for 2 h at 65 °C and subsequently digested with 3 units of DNase I (QIAGEN) for 2 h at 37 °C. After digestion, eluted samples were digested with proteinase K (0.75 mg/ml) for 2 h at 65 °C. RNA samples were purified using the RNeasy MinElute Cleanup kit (QIAGEN). Purified RNA samples were digested once more with 3 units of DNase I (QIAGEN) and purified. RNA samples were subjected to reverse-transcription before quantification by qPCR for different loci. RNA samples were split into 2 fractions. One fraction was used to measure the RNA levels and the other fraction was used as a negative control of reverse transcription (no reverse transcriptase added). The RNA was incubated at 90 °C for 1 min with 0.4 µl 25

mM dNTPs, 0.8 µl 5 µM of different primer pairs in 10 µl final volume. The RNA was then cooled down to 55 °C at a 0.8 °C/s temperature rate. A mix of 1 µl 100 mM DTT, 1 µl SuperScript III in 1x FS-buffer (Invitrogen) was added to the reactions. Negative control sample did not contain SuperScript III reverse-transcriptase. The RNA was reverse transcribed for 60 min at 55 °C. The enzyme was inactivated at 70 °C for 15 min. RNA samples were diluted with 30 µl H_2O and used in qPCR for quantification.

RNA quantification by RT-qPCR

Exponentially growing cells were collected and resuspended in 400 µl AE Buffer (50 mM sodium citrate in 10 mM EDTA pH5.3) and lysed with 500 µl calibrated phenol (with AE buffer) at 65 °C for 5 min. The aqueous phase was separated by centrifugation and mixed with 500 µl phenol-chloroform-isoamyl alcohol for 5 min at room temperature. The aqueous phase was again separated by centrifugation and collected. RNA was precipitated with 40 µl 3M sodium acetate and 1 ml 100% ethanol. RNA was pelleted by centrifugation and washed with 80% ethanol. Air-dried pellet was subsequently resuspended in a solution containing 3 µl DNase I (QIAGEN) in RDD buffer to digest genomic DNA. DNA was digested for 45 min at 37 °C and remaining RNA was purified with RNeasy MinElute Cleanup kit (QIAGEN). The RNA was incubated at 90 °C for 1 min with 0.4 µl 25 mM dNTPs, 0.8 µl 5 µM of different primer pairs in 10 µl final volume reaction. The RNA was then cooled-down to 55 °C at a 0.8 °C/s temperature rate. A mix of 1 µl 100 mM DTT, 1 µl SuperScript III in 1x FS-buffer (Invitrogen) was added to the reactions. Negative control sample did not contain SuperScript III reverse-transcriptase. The RNA was reverse transcribed for 60 min at 55 °C. The enzyme was inactivated at 70 °C for 15 min. RNA samples were diluted with 30 µl H_2O and subjected to qPCR. Data were processed according to the 2^(-ΔΔCt) method and expressed as %input. For both +/-RT conditions, %input was calculated normalized to the +RT input (5%) values. Finally, +RT %input were

normalized to the respective –RT %input. Error bars show SEM, calculated as $SEM=(SEM_1\text{^}2 + SEM_2\text{^}2)^{1/2}$.

Yeast transformation

Exponentially growing BY4742 yeast cells in YPD medium were pelleted (300g, 5 min) and gently resuspended in 3 ml SORB buffer (100 mM LiOAc, 10 mM Tris-HCl pH 8.0, 1 mM EDTA pH 8.0, 1 M Sorbitol). 50 µg of sheared salmon sperm (Ambion) carrier DNA was boiled 5 min at 95 °C prior to the addition of 100 ng of plasmid DNA. Cells were resuspended in 100 µL LiT solution (100 mM LiAc, 10 mM Tris-HCl pH 7.4), plasmid-carrier DNA was added and followed by the addition of 500 µL of PEG/LiT (polyethylene glycol 3350 (Sigma-Aldrich)). Samples were vortexed and incubated in a rotation wheel for 30 min at room temperature. 50 µL of DMSO (Sigma-Aldrich) were added and incubated for 15 min at 42 °C. Cells were pelleted (500 g for 30 sec) and resuspended in 200 µL of SD-URA medium, incubated at 30 °C for 30 min and plated on SD-URA plates for 2-3 days.

Flow cytometry

The relevant RNA folds were cloned into the corresponding UTR location of a centromeric GFP reporter plasmid. The resulting plasmids were used in the respective yeast knockout strains (Dharmacon). The BY4741 yeast knockout strains transformed with the GFP reporter plasmid were grown to saturation in SD-URA selection media at 30 °C, diluted 1:100 and further grown to OD_{600nm} of 0.7 - 0.9. Cells were analyzed by flow cytometry on a BD LSRFortessa SORP (BD Biosciences). Doublets were excluded via SSC-W signal and dead cells were excluded by DAPI staining. 20,000 events were measured per experiment and median values used for data analysis with FlowJo software (v10.5.3). GFP mean fluorescence intensities for the knockout experiments were normalized to the corresponding wildtype condition and the mean value of three experimental replicates was plotted. Error bars represent the standard deviation of the GFP knockout/wildtype values.

Pulsed SILAC

SILAC medium (6.7 g YNB w/o amino acids and with ammonium sulfate, 20 g Dextrose, 0.2 g L-adenine sulphate, 0.1 g L-tyrosine, 0.01 g L-histidine, 0.06 g L-leucine, 0.01 g L-methionine, 0.06 g L-phenylalanine, 0.04 g L-tryptophane, 0.02 g uracil and 0.02 g L-arginine per liter) was supplemented with 30 mg L^{-1} either lysine-0, lysine-4 and lysine-8 (Eurisotop) and sterile filtered through a 0.22 µm filter (Fisher Science). A preculture of 5 ml in lysine-0 medium was grown overnight at 30 ˚C. The culture was diluted and grown to OD_{600}=0.4 in lysine-0 medium, prior to two washes with PBS (cells pelleted by centrifugation at 500 g). When cells were resuspended in 15 ml lysine-8 medium (knockout strain) and lysine-4 medium (wildtype strain), actinomycin D was added at a final concentration of 1 µg mL^{-1}. Cells were grown with gentle agitation for 2 hours at 30 ˚C. Cells from both cultures were mixed at OD_{600}=0.5 and harvested by centrifugation at 14,000 g for 2 min at 4 ˚C. The pellet was washed with PBS and transferred to a clean tube. The cells were again centrifuged at 14,000 g for 15 s at 4 ˚C and resupended in 50 µL 1x NuPAGE LDS buffer (Thermo), sonicated for 10 cycles (30 sec ON/OFF), spun down at 14.000 g and 20 µL loaded in a 10% NuPage NOVEX precast gel (Thermo). Subsequent protein separation, staining and in-gel digest was done as described (see MS sample preparation and measurement section). Data was analyzed with MaxQuant and further processed with in-house scripts (see MS data analysis).

DATA AVAILABILITY

All relevant data are available from the authors. The source data underlying Figs 3b-c, 4b, 4e are provided as a Source Data file.

The mass spectrometry raw data is available at ProteomeXchange (http://www.proteomeexchange.org) under the data set identifier PXD014092 (Reviewer account details: Username: reviewer00504@ebi.ac.uk Password: nY5XhIi9).

CODE AVAILABILITY

Custom code used to analyze data in this study is available at https://github.com/ssayols/rnafold_interactome_casas_vila_et_al.

REFERENCES

1. Gerstberger, S., Hafner, M. & Tuschl, T. A census of human RNA-binding proteins. *Nat. Rev. Genet.* **15**, 829–845 (2014).
2. Ramanathan, M., Porter, D. F. & Khavari, P. A. Methods to study RNA–protein interactions. *Nature Methods* vol. 16 225–234 (2019).
3. Mitchell, S. F., Jain, S., She, M. & Parker, R. Global analysis of yeast mRNPs. *Nat. Struct. Mol. Biol.* **20**, 127–133 (2012).
4. Beckmann, B. M. *et al.* The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. *Nat. Commun.* **6**, 10127 (2015).
5. Matia-gonzález, A. M., Laing, E. E. & Gerber, A. P. Conserved mRNA-binding proteomes in eukaryotic organisms. *Nat. Publ. Gr.* **22**, 1027–1033 (2015).
6. Kwok, C. K., Tang, Y., Assmann, S. M. & Bevilacqua, P. C. The RNA structurome: Transcriptome-wide structure probing with next-generation sequencing. *Trends in Biochemical Sciences* vol. 40 221–232 (2015).
7. Rouskin, S., Zubradt, M., Washietl, S., Kellis, M. & Weissman, J. S. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature* **505**, 701–5 (2014).
8. Butter, F., Scheibe, M., Morl, M. & Mann, M. Unbiased RNA-protein interaction screen by quantitative proteomics. *Proc. Natl. Acad. Sci.* **106**, 10626–10631 (2009).
9. Srisawat, C. & Engelke, D. R. RNA affinity tags for purification of RNAs and ribonucleoprotein complexes. *Methods* (2002) doi:10.1016/S1046-2023(02)00018-X.

10. Castello, A. *et al.* Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. *Cell* **149**, 1393–1406 (2012).
11. Ma, W. K., Cloutier, S. C. & Tran, E. J. The DEAD-box protein Dbp2 functions with the RNA-binding protein Yra1 to promote mRNP assembly. *J. Mol. Biol.* **425**, 3824–3838 (2013).
12. Kikuma, T. *et al.* Dbp9p, a member of the DEAD box protein family, exhibits DNA helicase activity. *J. Biol. Chem.* **279**, 20692–20698 (2004).
13. Simpson, C. E., Lui, J., Kershaw, C. J., Sims, P. F. G. & Ashe, M. P. mRNA localization to Pbodies in yeast is biphasic with many mRNAs captured in a late Bfr1pdependent wave. *J. Cell Sci.* **127**, 1254–1262 (2014).
14. Weaver, P. L., Sun, C. & Chang, T. H. Dbp3p, a putative RNA helicase in Saccharomyces cerevisiae, is required for efficient pre-rRNA processing predominantly at site A3. *Mol. Cell. Biol.* **17**, 1354–65 (1997).
15. Frugier, M. & Giegé, R. Yeast aspartyl-tRNA synthetase binds specifically its own mRNA. *J. Mol. Biol.* **331**, 375–383 (2003).
16. Sampath, P. *et al.* Noncanonical function of glutamyl-prolyl-tRNA synthetase: Gene-specific silencing of translation. *Cell* **119**, 195–208 (2004).
17. Suliman, H. S., Sawyer, G. M., Appling, D. R. & Robertus, J. D. Purification and properties of cobalamin-independent methionine synthase from Candida albicans and Saccharomyces cerevisiae. *Arch. Biochem. Biophys.* **441**, 56–63 (2005).
18. Kingsbury, J. M. & McCusker, J. H. Cytocidal amino acid starvation of Saccharomyces cerevisiae and Candida albicans acetolactate synthase (ilv2Δ) mutants is influenced by the carbon source and rapamycin. *Microbiology* **156**, 929–939 (2010).
19. Crabeel, M., Lavalle, R. & Glansdorff, N. Arginine-specific repression in Saccharomyces cerevisiae: kinetic data on ARG1 and ARG3 mRNA transcription and stability support a transcriptional control mechanism. *Mol. Cell. Biol.* **10**, 1226–1233 (2015).

20. Alifano, P. et al. Histidine Biosynthetic Pathway and Genes: Structure, Regulation, and Evolution. MICROBIOLOGICAL REVIEWS vol. 60 (1996).
21. Galani K, Grosshands H, Deinert K, C.Hurt E & Simos G. The intracellular location of two aminoacyl-tRNA synthetases depends on complex formation with Arc1p. *EMBO J.* **20**, 6889–6898 (2001).
22. Liu, Q., Krzewska, J., Liberek, K. & Craig, E. A. Mitochondrial Hsp70 Ssc1: Role in Protein Folding. *J. Biol. Chem.* **276**, 6112–6118 (2001).
23. Abrams, J. L., Verghese, J., Gibney, P. A. & Morano, K. A. Hierarchical functional specificity of cytosolic heat shock protein 70 (Hsp70) nucleotide exchange factors in yeast. *J. Biol. Chem.* **289**, 13155–13167 (2014).
24. Umemotos, N., Yoshihisa$, T., Hiratas, R. & Anrakusq, Y. Gene Product, Subunit c of the Vacuolar Membrane H'-ATPase on Vacuolar Acidification and Protein Transport A STUDY WITH VMA3-DISRUPTED MUTANTS OF SACCHAROMYCES CEREVISIAE*. THE JOURNAL OF BIOLOGICAL CHEMISTRY vol. 265 http://www.jbc.org/ (1990).
25. Walther, T. C. *et al.* Eisosomes mark static sites of endocytosis. *Nature* **439**, 998–1003 (2006).
26. Calabretta, S. & Richard, S. Emerging Roles of Disordered Sequences in RNA-Binding Proteins. *Trends in Biochemical Sciences* (2015) doi:10.1016/j.tibs.2015.08.012.
27. Fleischer, T. C., Weaver, C. M., McAfee, K. J., Jennings, J. L. & Link, A. J. Systematic identification and functional screens of uncharacterized proteins associated with eukaryotic ribosomal complexes. *Genes Dev.* **20**, 1294–1307 (2006).
28. Mazan-Mamczarz, K. *et al.* Targeted suppression of MCT-1 attenuates the malignant phenotype through a translational mechanism. *Leuk. Res.* **33**, 474–482 (2009).
29. Aravind, L. & Koonin, E. V. Novel Predicted RNA-Binding Domains Associated with the Translation Machinery.

30. Reinert, L. S. *et al.* MCT-1 protein interacts with the cap complex and modulates messenger RNA translational profiles. *Cancer Res.* **66**, 8994–9001 (2006).
31. Kwon, S. C. *et al.* The RNA-binding protein repertoire of embryonic stem cells. *Nat. Struct. Mol. Biol.* (2013) doi:10.1038/nsmb.2638.
32. Zhu, Y. *et al.* POSTAR2: Deciphering the post-Transcriptional regulatory logics. *Nucleic Acids Res.* **47**, D203–D211 (2019).
33. Costanzo, M. *et al.* A global genetic interaction network maps a wiring diagram of cellular function. doi:10.1126/science.
34. Dmitriev, S. E. *et al.* GTP-independent tRNA delivery to the ribosomal P-site by a novel eukaryotic translation factor. *J. Biol. Chem.* **285**, 26779–26787 (2010).
35. Krogan, N. J. *et al.* High-Definition Macromolecular Composition of Yeast RNA-Processing Complexes. *Mol. Cell* **13**, 225–239 (2004).
36. Measday, V. *et al.* A family of cyclin-like proteins that interact with the Pho85 cyclin-dependent kinase. *Mol. Cell. Biol.* **17**, 1212–1223 (2015).
37. O'Neill, E., Kaffman, A., Jolly, E. & O'Shea, E. Regulation of PHO4 Nuclear Localization by the PHO80-PHO85 Cyclin-CDK Comlex. *Science (80-.).* **271**, 209–212 (1996).
38. Shemer, R., Meimoun, A., Holtzman, T. & Kornitzer, D. Regulation of the Transcription Factor Gcn4 by Pho85 Cyclin Pcl5. *Mol. Cell. Biol.* **22**, 5395–5404 (2002).
39. Lee, W. C., Zabetakis, D. & Mélèse, T. NSR1 is required for pre-rRNA processing and for the proper maintenance of steady-state levels of ribosomal subunits. *Mol. Cell. Biol.* **12**, 3865–3871 (2015).
40. Gerber, A. P., Herschlag, D. & Brown, P. O. Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. *PLoS Biol.* (2004) doi:10.1371/journal.pbio.0020079.
41. Miller, M. a & Olivas, W. M. Roles of Puf proteins in mRNA degradation and translation. *Wiley Interdiscip. Rev. RNA* **2**, 471–92 (2011).

42. Webster, M. W., Stowell, J. A. & Passmore, L. A. RNA-binding proteins distinguish between similar sequence motifs to promote targeted deadenylation by Ccr4-Not. *Elife* **8**, (2019).
43. Tarassov, K. *et al. An in Vivo Map of the Yeast Protein Interactome*. http://science.sciencemag.org/ (2006).
44. Senissar, M. *et al.* The DEAD-box helicase Ded1 from yeast is an mRNP cap-associated protein that shuttles between the cytoplasm and nucleus. *Nucleic Acids Res.* **42**, 10005–10022 (2014).
45. Noueiry, A. O., Chen, J. & Ahlquist, P. A mutant allele of essential, general translation initiation factor DED1 selectively inhibits translation of a viral mRNA. *Proc. Natl. Acad. Sci.* **97**, 12985–12990 (2002).
46. Amrani, N., Minet, M., Le Gouar, M., Lacroute, F. & Wyers, F. Yeast Pab1 interacts with Rna15 and participates in the control of the poly(A) tail length in vitro. *Mol. Cell. Biol.* **17**, 3694–701 (1997).
47. Thiel, B. C., Ochsenreiter, R., Gadekar, V. P., Tanzer, A. & Hofacker, I. L. RNA structure elements conserved between mouse and 59 other vertebrates. *Genes (Basel).* **9**, (2018).
48. Thompson, J. D., Higgins+, D. G. & Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research vol. 22 (1994).
49. Lorenz, R. *et al. ViennaRNA Package 2.0*. http://www.tbi.univie.ac.at/RNA. (2011).
50. Bunn, A. G. A dendrochronology program library in R (dplR). *Dendrochronologia* **26**, 115–124 (2008).
51. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12**, 115–121 (2015).
52. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).

53. Olsen, J. V. *et al.* Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **4**, 709–712 (2007).
54. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–72 (2008).
55. Cox, J. *et al.* Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).
56. Cherry, J. M. *et al.* Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* **40**, D700-5 (2012).
57. Mészáros, B., Erdös, G. & Dosztányi, Z. IUPred2A: Context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **46**, W329–W337 (2018).
58. Csárdi, G. & Nepusz, T. The igraph software package for complex network research. (2006).
59. Baejen, C. *et al.* Transcriptome maps of mRNP biogenesis factors define pre-mRNA recognition. *Mol Cell* **55**, 745–757 (2014).
60. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
61. Uren, P. J. *et al.* Site identification in high-throughput RNA-protein interaction data. *Bioinformatics* **28**, 3013–3020 (2012).

ACKNOWLEDGEMENTS

We were supported by the Proteomics Core Facility, the Cytometry Core Facility and the Media Lab at IMB. We thank ... and the Weissman lab for helpful discussion. Work was funded by the Deutsche Forschungsgemeinschaft as part of the Priority Programme SPP1935 ("Deciphering the mRNP code: RNA-bound Determinants of Post-transcriptional Gene Regulation) to F.B. (Bu 2996/5-1).

AUTHOR CONTRIBUTIONS

N.C., M.S. and F.B. conceived the study; N.C., L.P. and M.S. performed the experiments; N.C. and S.S. did the bioinformatics analysis; N.C. and F.B. wrote the paper with input from all authors.

COMPETING INTERESTS

Authors declare no competing financial interests.

TABLES

Table 1. Motif analysis on Puf3 targets.

mRNA	UGUAAAUA	Motif location	Similar motifs
COX13		3'UTR	UGUAAA
COX9			
MRPL35	UGUAAAUA	CDS	
ATP12	UGUAAAUA	3'UTR	
ATP7		3'UTR	GUAAAUA
ATP17	UGUAAAUA	5'UTR	
ATP20		3'UTR	GUAAAUA
ATP4			
MZM1	UGUAAAUA	3'UTR	
COQ1	UGUAAAUA	3'UTR	
MIR1		3'UTR	UGUAAAU
CMC2	UGUAAAUA	5'UTR/3'UTR	

SUPPLEMENTARY DATA

Supplementary Data 1. List of protein interactors identified by MaxQuant analysis.

Supplementary Data 2. List of mRNA-protein interactors in the heatmap of Figure 1d, showing the enrichment value for each scored mRNA-protein pair.

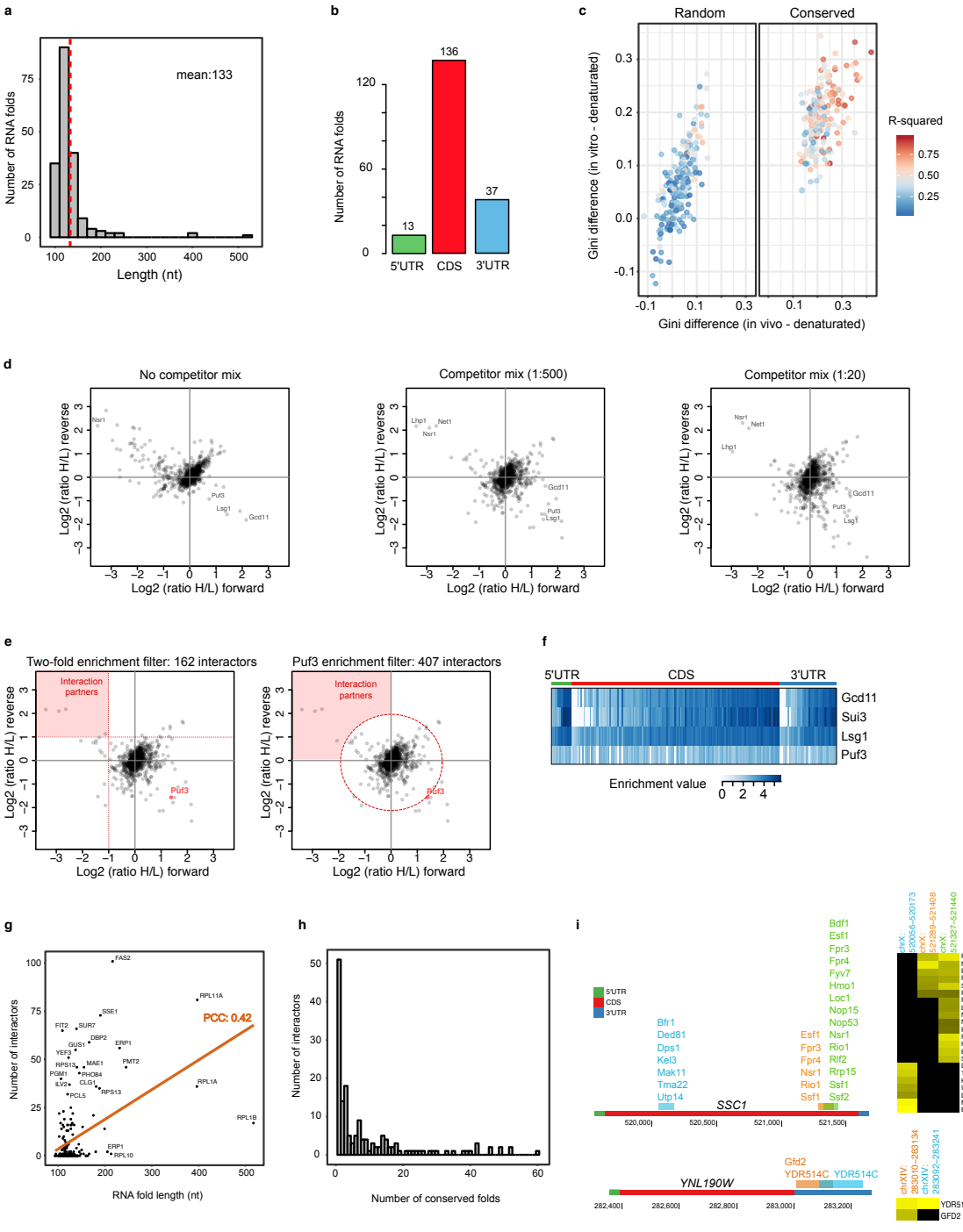
Supplementary Data 3. List of GO terms and RNA-binding domains used in our interactor set analysis.

Supplementary Data 4. GO term enrichment analysis of similar genetic interaction profile communities of our protein interactors and their mRNA target genes.

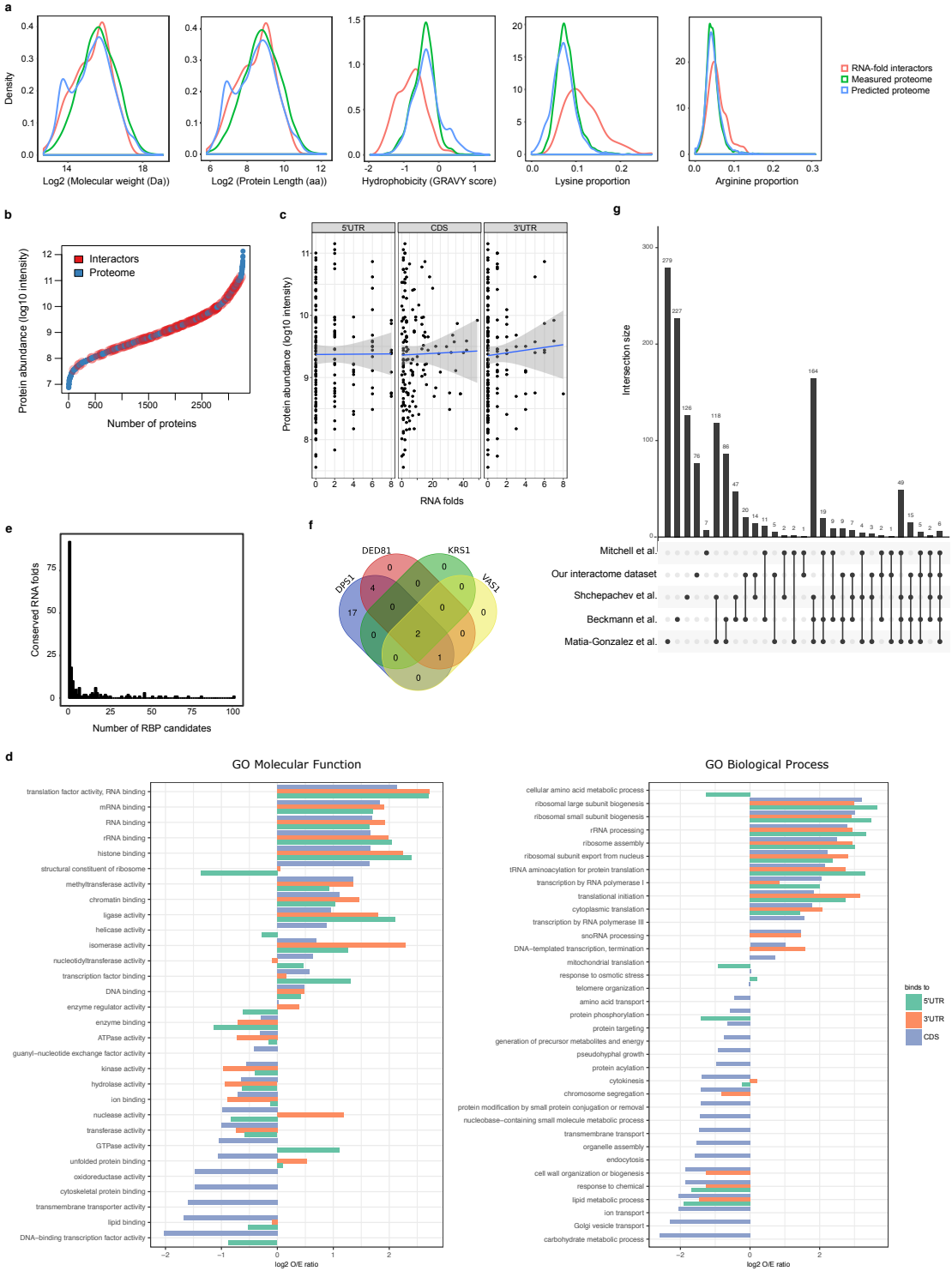
Supplementary Data 5. List of primers and yeast strains used in this study.

Supplementary Data 6. 162 interactors ranked by fold type and GO term enrichment analysis on 5'UTR, CDS and 3'UTR interactors.

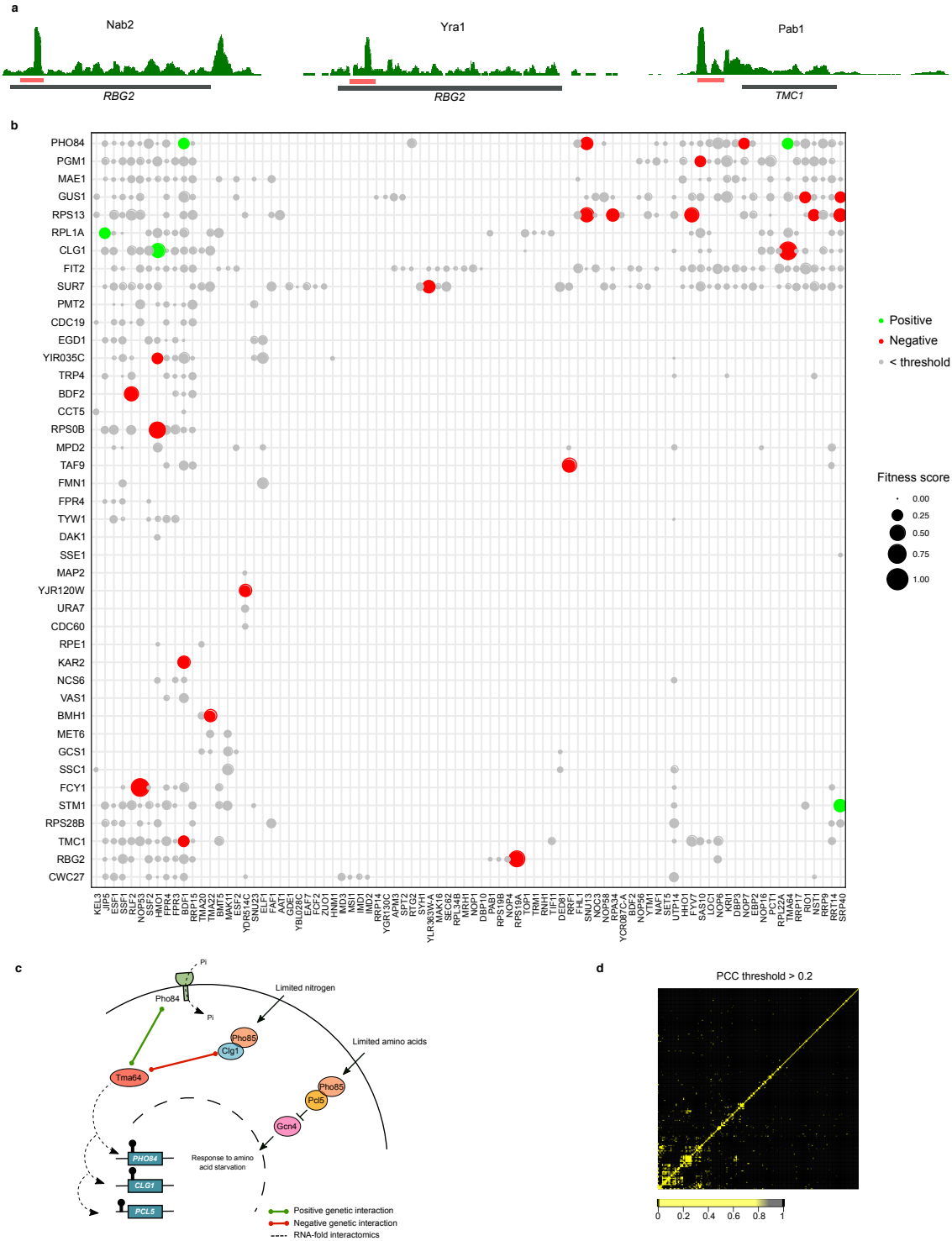
Supplementary Data 7. Pulsed SILAC results for *puf3*, *nsr1* and *tma20*.


a
mean:133
Number of RNA folds
Length (nt)
b
Number of RNA folds
136
37
13
5'UTR
CDS
3'UTR
c
Random
Conserved
Gini difference (in vitro - denaturated)
Gini difference (in vivo - denaturated)
R-squared
d
No competitor mix
Competitor mix (1:500)
Competitor mix (1:20)
Log2 (ratio H/L) reverse
Log2 (ratio H/L) forward
e
Two-fold enrichment filter: 162 interactors
Puf3 enrichment filter: 407 interactors
Interaction partners
f
5'UTR
CDS
3'UTR
Gcd11
Sui3
Lsg1
Puf3
Enrichment value
g
PCC: 0.42
Number of interactors
RNA fold length (nt)
h
Number of interactors
Number of conserved folds
i
5'UTR
CDS
3'UTR
SSC1
YNL190W
Gfd2
YDR514C
Enrichment value

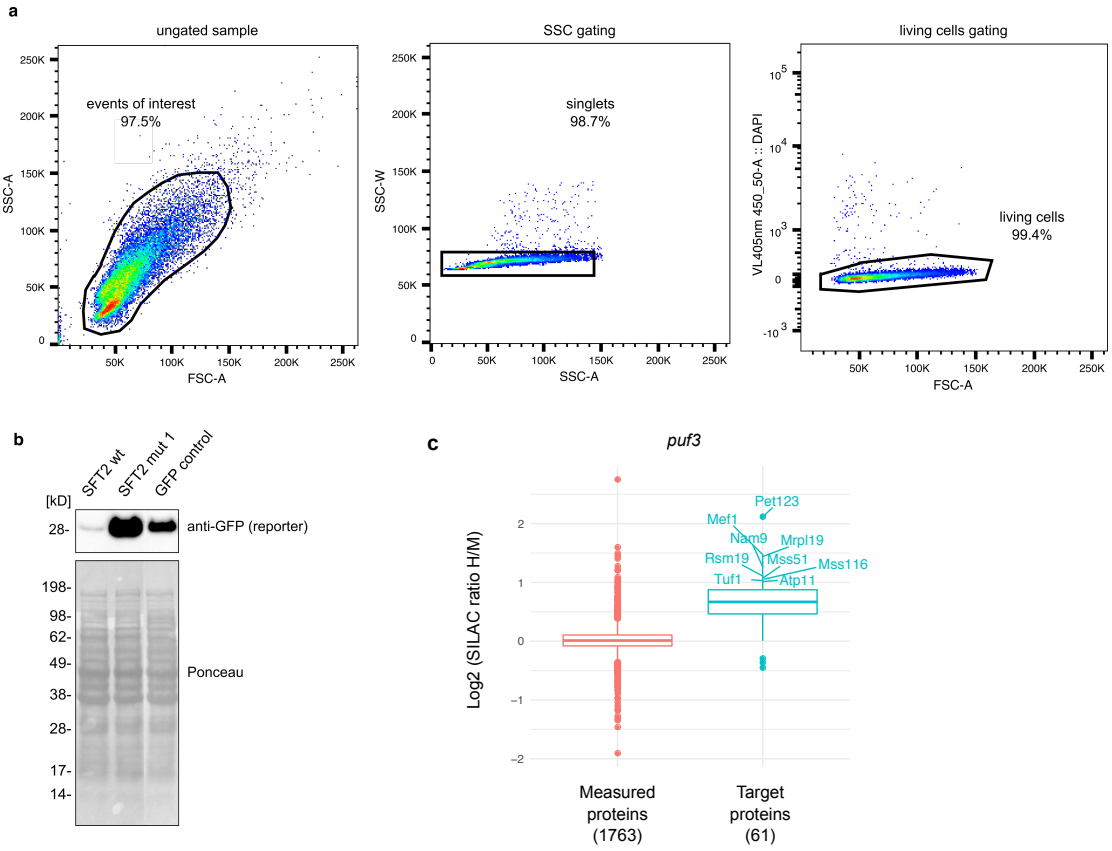
Supplementary Figure 1. RNA fold features and technical aspects of our RNA fold interactome.
(a) Length (nt) distribution of the 188 evolutionary conserved RNA folds. Red-dashed line indicates mean value.
(b) Distribution of the conserved RNA folds according to location (5'UTR, CDS, 3'UTR).
(c) Structure analysis based on *in vitro* and *in vivo* DMS-Seq data comparing 188 random genomic regions to the set of 188 evolutionary conserved RNA structures. Folds are colored according to *in vitro*/*in vivo* DMS-Seq correlations. R-squared coloring according to the degree of correlation between *in vivo* and *in vitro* conditions.
(d) Two-dimensional interaction plot showing protein enrichment for one RNA fold using different competitor mix concentrations (0, 1:500 and 1:20 dilutions).
(e) Examples of two different filtering criteria to score for interaction partners. Red area indicates interaction partners.
(f) Enrichment values for the four control proteins Puf3, Gcd11, Sui3 and Lsg1 at the *COX17* UTR in all 186 pull-down experiments.
(g) Correlation plot on number of RBP candidates binding one structure vs. structure length. PCC: Pearson Correlation Coefficient.
(h) Histogram showing the number of RBP interactors binding to each RNA fold.
(i) Schematics of RNA fold localization on the *SSC1* and *YNL190W* mRNAs harboring multiple folds and the corresponding interactors. Enrichment values for each interaction are shown in the heatmap.


a
Density
Log2 (Molecular weight (Da))
Log2 (Protein Length (aa))
Hydrophobicity (GRAVY score)
Lysine proportion
Arginine proportion
RNA-fold interactors
Measured proteome
Predicted proteome
b
Interactors
Proteome
Protein abundance (log10 intensity)
Number of proteins
c
5'UTR
CDS
3'UTR
Protein abundance (log10 intensity)
RNA folds
g
Intersection size
Mitchell et al.
Our interactome dataset
Shchepachev et al.
Beckmann et al.
Matia-Gonzalez et al.
e
Conserved RNA folds
Number of RBP candidates
f
DED81
KRS1
DPS1
VAS1
d
GO Molecular Function
translation factor activity, RNA binding
mRNA binding
RNA binding
rRNA binding
histone binding
structural constituent of ribosome
methyltransferase activity
chromatin binding
ligase activity
helicase activity
isomerase activity
nucleotidyltransferase activity
transcription factor binding
DNA binding
enzyme regulator activity
enzyme binding
ATPase activity
guanyl−nucleotide exchange factor activity
kinase activity
hydrolase activity
ion binding
nuclease activity
transferase activity
GTPase activity
unfolded protein binding
oxidoreductase activity
cytoskeletal protein binding
transmembrane transporter activity
lipid binding
DNA−binding transcription factor activity
log2 O/E ratio
GO Biological Process
cellular amino acid metabolic process
ribosomal large subunit biogenesis
ribosomal small subunit biogenesis
rRNA processing
ribosome assembly
ribosomal subunit export from nucleus
tRNA aminoacylation for protein translation
transcription by RNA polymerase I
translational initiation
cytoplasmic translation
transcription by RNA polymerase III
snoRNA processing
DNA−templated transcription, termination
mitochondrial translation
response to osmotic stress
telomere organization
amino acid transport
protein phosphorylation
protein targeting
generation of precursor metabolites and energy
pseudohyphal growth
protein acylation
cytokinesis
chromosome segregation
protein modification by small protein conjugation or removal
nucleobase−containing small molecule metabolic process
transmembrane transport
organelle assembly
endocytosis
cell wall organization or biogenesis
response to chemical
lipid metabolic process
ion transport
Golgi vesicle transport
carbohydrate metabolic process
log2 O/E ratio
binds to
5'UTR
3'UTR
CDS

Supplementary Figure 2. Characterization of our RNA fold interactors and comparison to other studies.
(a) Density plots showing biochemical features of our RNA fold interactome (orange), measured proteome (green) and predicted proteome (blue).
(b) Ranked protein abundance plotted against enriched interactors reveals that interactors (red) cover the full proteome abundance range.
(c) Protein abundance grouped by fold type (5'UTR, CDS, 3'UTR). Shaded area represents confidence interval (0.95) around the linear model.
(d) GO term enrichment analysis on molecular function and biological process of interactors differentiated by location within the target RNA (5'UTR, CDS, 3'UTR). The log2 values of the observed/expected ratio for the top25 GO terms of ranked by fold type are shown.
(e) Histogram showing the number of conserved folds bound by each RBP candidate.
(f) Overlap of target genes for the four identified tRNA synthetases.
(g) Comparison of our RBP candidates with other existing interactome capture studies in *S. cerevisiae*.


a
Nab2
RBG2
Yra1
RBG2
Pab1
TMC1
b
PHO84
PGM1
MAE1
GUS1
RPS13
RPL1A
CLG1
FIT2
SUR7
PMT2
CDC19
EGD1
YIR035C
TRP4
BDF2
CCT5
RPS0B
MPD2
TAF9
FMN1
FPR4
TYW1
DAK1
SSE1
MAP2
YJR120W
URA7
CDC60
RPE1
KAR2
NCS6
VAS1
BMH1
MET6
GCS1
SSC1
FCY1
STM1
RPS28B
TMC1
RBG2
CWC27
Positive
Negative
< threshold
Fitness score
0.00
0.25
0.50
0.75
1.00
c
Pi
Pho84
Pi
Limited nitrogen
Pho85
Clg1
Limited amino acids
Tma64
Pho85
Pcl5
Gcn4
PHO84
CLG1
PCL5
Response to amino acid starvation
Positive genetic interaction
Negative genetic interaction
RNA-fold interactomics
d
PCC threshold > 0.2
0
0.2
0.4
0.6
0.8
1

Supplementary Figure 3. RNA fold-protein interaction validation and genetic interaction data integration.
(a) PAR-CLIP data for Pab1, Yra1 and Nab2 validates binding to the investigated RNA folds. The location of the corresponding evolutionary conserved RNA fold is indicated in salmon.
(b) Complete matrix of genetic interactions described for our RBP (x-axis) and RNA fold (y-axis) interacting pairs. Genetic interactions with a fitness score > 0.08 are colored according to positive (green) and negative (red) interactions and the remaining are shown in grey. The circle size is proportional to the fitness score of the double knockout strain of the two relevant genes.
(c) Integration of literature knowledge of Tma64 and functional evidences based on our interactomics screen and genetic interaction data.
(d) Distance correlation of genetic interaction profiles of all genes (interactors and mRNAs). Correlations higher than 0.2 are highlighted in yellow and filtered for similar genetic interaction profiles.



Supplementary Figure 4. Technical aspects of the reporter screen and positive control for pulsed SILAC.
(a) Example of gating used for GFP fluorescence intensities quantification by flow cytometry.
(b) Immunostaining shows changes in GFP levels upon fusion of the wildtype or the mutant 5'UTR *SFT2* loop as previously validated (Rouskin et al., 2014), and examined in two independent biological experiments.
(c) Pulsed SILAC box plots for *puf3* knockout compared to a wildtype strain. Known target genes (blue) are upregulated compared to all measured proteins (salmon) (A. P. Gerber, Herschlag, & Brown, 2004). Boxes show median (center) and interquartile ranges (ends), lower whisker representing the smallest observation greater than or equal to 1.5 times the interquartile range and upper whisker representing the largest observation less than or equal to 1.5 times the interquartile range.

V. Discussion and Outlook

DISCUSSION

Quantitative proteomics to study organismal development

In the first study we generated two large proteomic datasets covering embryogenesis and the whole life cycle of *Drosophila melanogaster* using label-free quantitative mass spectrometry. Previous proteomic studies were restricted to specific developmental stages like embryo or the oocyte-to-embryo transition (Fabre et al., 2016; Kronja et al., 2014), adult flies (Sury et al., 2010) or larva and pupa stages (Chang et al., 2013; Xing et al., 2014). Other studies have used adapted cultured lines or specific proteins for which antibodies or tagged lines are available to investigate protein dynamics at certain stages (Brunner et al., 2007; Nagarkar-Jaiswal et al., 2015; Sarov et al., 2016). Our whole life cycle dataset for the whole life cycle includes 15 time-points of embryo, larva, pupa and adult fly stages and the second dataset contains 14 time points with highly temporal resolution during embryogenesis. As a result, we here extend the current proteomics data to nearly 8,000 proteins during the life cycle and 5,500 proteins in the embryogenesis process, accounting for almost half of the annotated *Drosophila* genes (17,869, FlyBase FB2020_01 release) and nearly two thirds of reported protein coding genes (13,968, FlyBase FB2020_01 release). Because our study includes different stages and time-points, we are able to score 1,535 developmentally regulated proteins during the whole life cycle and 1,644 during embryogenesis.

Almost half of the proteins detected during the life cycle are functionally not known. Therefore, our dataset reveals relevant temporal windows to study phenotypic effects of yet uncharacterized proteins. Integration of tissue-specific expression data adds a spatial dimension for functional studies. mRNA location correlates well with protein function. Hence, tissue expression data from a large fluorescence *in situ* RNA hybridization study was combined with our data to infer tissue-specific protein functions (Lécuyer et al., 2007; Wilk, Hu, Blotsky, & Krause, 2016). As a result, we characterized a role for CG1674 in muscle development during embryogenesis. This example underscores the value of our

dataset in combination with tissue expression data and provides a powerful resource for future studies.

Drosophila development has widely been addressed by transcriptome studies using microarray or RNA-sequencing datasets (Brown et al., 2014; Tomancak et al., 2007; Graveley et al., 2011). However, RNA levels cannot be used as a proxy for protein abundance since moderate correlation between transcriptome and proteome has been noted in several model organisms such as *S. cerevisiae* (Griffin et al., 2002), *C. elegans* (Grün et al., 2014), *X. laevis* (Peshkin et al., 2015) and humans (Y. Liu, Beyer, & Aebersold, 2016). In this study, we compared our embryogenesis dataset to the modENCODE transcriptome data generated by Graveley et al. and noted limited mRNA/protein correlation as well. This can be partially attributed to temporally delayed protein expression relative to mRNA. In fact, the best correlation is nonsynchronous, with an overall 4- to 5-hours proteome lag, probably due to protein synthesis delay. Protein-mRNA correlations have been shown to improve in later development time-points, possibly due to less pronounced dynamical changes. However, even if time delays are taken into account, they still don't reach complete correlation in later time-points (Becker et al., 2018). Protein abundance can also be regulated at a translational level. For instance, the ubiquitin-proteasome degradation system controls protein degradation rates independently of transcript levels (Lau et al., 2018; McShane et al., 2016). Additionally, the global mismatch between protein and mRNA levels can also be attributed to post-transcriptional gene regulation mechanisms in eukaryotic organisms. Global studies in mammalian cells and yeast have reported that the expression of 20-30% of genes is controlled post-transcriptionally (Robles et al., 2014; Ingolia et al., 2009). Previous studies show that protein translation rates vary across mRNAs (Schwanhüusser et al., 2011), therefore ribosomal occupancy has been postulated as a better predictor of protein concentrations than mRNA concentrations (Ingolia et al., 2009). Discrepancy between protein-mRNA is even more pronounced in dynamical cellular transitions, where active regulation by RBPs, micro-RNAs or RNA modifications is enforced (Glisovic et al., 2008; Filipowicz et al., 2008; Roignant

& Soller, 2017). Thus, it is not surprising that the mRNA-bound proteome is highly dynamic during embryonic development (Schwanhüusser et al., 2011).
Early embryogenesis represents an interesting case to study post-transcriptional regulation because zygotic transcription in not switched on. During the first two-three hours of Drosophila embryogenesis, mRNA and protein dynamics rely on post-transcriptional regulation of maternal material until the maternal-to-zygotic transition (MZT) has taken place (Tadros & Lipshitz, 2009). Some maternally deposited mRNAs essential for early embryonic development such as *nanos* are post-transcriptionally regulated (Andrews, Snowflack, Clark, & Gavis, 2011). Important players in such processes are RNA-binding proteins like Smaug that control maternal transcript destruction (Benoit et al., 2009). We systematically identified maternally loaded proteins and reported a catalogue of not yet characterized proteins with possible functions in early development. The functional importance is exemplified by CG17018 knockdown in our study, provided that loss of CG17018 leads to non-viable embryos. Also, in a follow-up study by Becker et al., we generated highly time-resolved paired transcriptome/proteome data that allowed us to investigate post-transcriptionally regulated proteins during embryogenesis. Notably, we observed differences in developmental speed between our paired RNA-seq data and the previous published dataset by Gravely et al., underscoring the impact of laboratory conditions even for a robust embryonic developmental process. Mathematical models in the Becker et al. study classified most mRNA-protein pairs (84%) into four distinct regulatory scenarios taking translation and degradation into account. From in-depth analysis on the group of potentially post-transcriptionally regulated genes, we found significant motif enrichment of the RNA-binding protein Hrb98DE. Its target genes are differentially spliced upon protein knockdown and are associated with glucose metabolism by GO terms analysis. Overall, suggesting a role for Hrb98DE in post-transcriptional regulation of glucose metabolism genes, possibly at the level of pre-mRNA splicing.

Alternative splicing (AS) greatly expands proteome diversity. Isoform expression patterns may differ during development and confer tissue specialized functions (Baralle & Giudice, 2017). A previous transcriptome study reported that 20-37% of multi-exon genes are alternatively spliced in Drosophila (Gibilisco, Zhou, Mahajan, & Bachtrog, 2016). However, the detection of alternatively spliced gene variants by proteomics is hindered by the fact that the majority of protein sequence is shared between splice variants. Therefore, peptides either originating from an isoform-specific exon or from an exon-exon junction are used for quantification. While most protein isoform evidences come from individual experiments, a large study unambiguously detected over a hundred genes with at least two isoforms in Drosophila (Tress, Bodenmiller, Aebersold, & Valencia, 2008). Among them is *lola*, one of the most spliced loci with stage- and tissue-specific functions in *Drosophila* (Dinges, Morin, Kreim, Southall, & Roignant, 2017). Our embryogenesis proteomics dataset allows protein isoform quantification with high developmental resolution. We quantified multiple isoforms in 34 genes, some of them showing differential isoform expression during embryogenesis. We detected an expression peak of one of *lola*'s isoforms that was validated by immunoblotting using an isoform-specific antibody and suggests a stage-specific isoform function at late embryogenesis.

Despite achieving large protein coverage, our datasets are incomplete. Lowly expressed proteins and/or proteins or isoforms limited to specific tissues might be missing in our study. Therefore, a non-quantified protein does not necessarily mean that it is absent at a specific stage but can also be expressed below our limit of detection (LOD). Comprehensive proteome measurement is challenged by sample complexity, limited material as well as not sensitive enough MS workflows and instruments to achieve complete sequence coverage. Gel fractionation, biochemical methods or longer HPLC separation gradients can be used to separate a complex protein mixture and increase protein identification and quantification. Alternatively, other methods than the Top10 used in this study can be applied to select more than 10 peptide ions for

fragmentation. However, these strategies considerably increase measurement time that may compromise measurement stability in large-scale studies and do not always result in a significant increase of peptide identifications. Furthermore, advances in low-input sample workflows that minimize sample loss and reduce manual intervention have been recently proposed (Kulak et al., 2014; Hughes et al., 2019; Müller et al., 2020).

In summary, our high temporally resolved proteomics study across Drosophila development provides a valuable resource for future studies on individual proteins or at a system level. For example, phenotypic effects of a protein of interest can be evaluated at a stage and time in which the protein is expressed. Dynamic protein expression profiles are functionally very informative. Hence our interactive visualization tool provides a search option based on proteins with similar expression profiles that can hint towards functional protein groups, as seen for ecdysone-related proteins. Comparative analysis also pinpoint uncharacterized proteins that may be relevant to specific processes. For instance, we provide a comprehensive list of sex-and age-specific proteins, as well as propose novel candidates with an essential role in early embryogenesis by being maternally deposited. Integration with tissue-specific expression data is a powerful resource towards the characterization of proteins with time- and tissue-restricted functions, as exemplified with CG1674. In a short-term perspective, our proteomics resource will help generate multiple follow-up projects on different research areas in Drosophila.

A next interesting perspective to address would be the developmental proteome with tissue-specific resolution. Different tissue markers could be used to sort specific cell types during development by flow cytometry coupled to MS-based proteomics. Alternatively, manual tissue dissection on the microscope could be used. Significant improvements in sample preparation protocols from tissue samples have been introduced by scRNA-seq workflows (Nguyen, Pervolarakis, Nee, & Kessenbrock, 2018). Additionally, methodological advances in the proteomics field enable high proteome coverage from low-input protocols. The SP3 method presented by Krijgsveld

and colleagues minimizes sample loss and allows robust quantification of 500-1,000 proteins from 100 to 1,000 cells in clinical samples. So, even in very early time-points when cell numbers are limiting or for low-prevalence tissue markers, the most abundant proteins could be quantified.

Additional regulatory mechanisms would be interesting to investigate from a developmental perspective. PTMs regulate protein function, stability, interactions, activity and localization. To date, a phosphoproteomics study lacking developmental resolution was performed with a bulk of 0-24 hours embryos (Nguyen et al., 2018). Proteins that are phosphorylated during egg activation are functionally important for the egg-to-embryo transition and maternal mRNAs clearance (Zhang et al., 2018; C. Liu et al., 2018). Alterations in the phosphoproteome targeting kinases and substrates like transcription factors will also increase our knowledge about signaling networks during development. In addition, changes in ubiquitinated proteins can give insights into the regulation of maternal components degradation during the MZT. To comprehensively quantify PTMs, different methods for enrichment of modified peptides need to be applied during the MS workflow.

Further parallel developments in MS instrumentation, liquid chromatography and analysis pipelines will surely continue to improve measurement speed, sensitivity and reliable protein quantification. In a long-term perspective, single-cell proteomics of an embryo can also be envisioned. While recent work showed quantification of hundreds of proteins from a single-cell by mass spectrometry, improvements towards protocols with minimal sample loss and higher instrument sensitivity are required. Deep coverage of different marker genes would be required to address the challenge of mapping each cell type to the correct spatial position. Information on protein temporal dynamics of thousands individual cells will enormously increase our knowledge about mechanisms that shape cell identify during embryogenesis.

RNA-fold interaction proteomics

In the second article of this thesis we applied SILAC-based quantitative proteomics to 186 evolutionary conserved RNA structures. As a result, we report binding of 162 RBPs with RNA-fold resolution. Compared to global RNA interactome capture methods, which globally describe RBPs bound to polyadenylated RNA, we are able to associate individual RNA folds within mRNA with their set of interacting proteins. In our SILAC setup, a common control RNA is used against all evaluated RNA structures. Alternatively, proteins binding to a certain structure can be evaluated against a mutated version of the same RNA fold. This strategy is challenging for complex structures with many stem-loop formations because mutations at each hairpin need to be evaluated separately against a wildtype version. We used this strategy for the *PMA1* RNA fold and discover Sbp1, a known translational repressor, as a specific binder for the wildtype structure.

It is possible that we are not able to detect the complete set of interactors of each evaluated RNA fold. A SILAC ratio cannot be calculated for proteins with missing values in either the light or heavy SILAC experiment. While missing values in proteomics workflows can be attributed to multiple reasons (miscleavage, dynamic range, peptide misidentification, ambiguous matching, etc.) (Lazar, Gatto, Ferro, Bruley, & Burger, 2016), they are more frequent for peptides whose abundances are close to the limit of detection of the instrument. Numerous imputation algorithms have been developed to tackle the missing value problem either at peptide- or protein-level (Webb-Robertson et al., 2015), and it is therefore important to carefully evaluate the effect of imputation on each proteomic dataset. The number of binding partners is surely also influenced by the experimental set-up. For instance, the stringency of the wash conditions during IP-MS can preserve weak interactions or only report high affinity interactions. Also, the use of cross-linking or high stringency washing permits removing proteins bound non-specifically (Scheibe et al., 2013). Importantly, our data shows no correlation between protein abundance and RNA fold binding, binding to multiple RNA folds or specific fold

types (5'UTR, CDS, 3'UTR) and therefore we are confident that we report physiologically relevant interactions.

While globally RNA appears more structured *in vitro* compared to *in vivo*, we show that our investigated set of conserved RNA folds present similar folding (Rouskin et al., 2014)(Figure 1 of Chapter IV). This sets us well to study meaningful interactions from the RNA structure perspective. We acknowledge the fact that some interactions are missed because our study is limited to fully recapitulate all possible cellular scenarios. Indeed, availability of competing proteins and RNAs at specific cellular contexts can modulate binding of RBPs, as well as changes on local concentrations of metabolites and ions in response to biological cues.

Whereas almost two thirds of our interactor set is enriched for RBP features, the remaining proteins are unrelated to RNA-binding and do not harbour canonical RBDs. It is important to consider that our setup does not only capture proteins in direct contact with RNA but also enriches for RBP complexes. Nevertheless, not all discovered RBPs harbor canonical RBDs, but also include proteins like kinases, metabolic enzymes and tRNA and rRNA-related proteins (Castello et al., 2015). Proteins with intrinsically disordered regions can also adopt specific conformations that allow interactions with RNA (Calabretta & Richard, 2015). Therefore, our data can shed light into mRNA-protein interactions via unconventional protein domains. In line with this idea, we report four tRNA synthetases characterized by a disordered N-terminal extension in contrast to other yeast tRNA synthetases. While a role in translation regulation for the tRNA synthetases Dps1 has been reported, it will be exciting to evaluate whether this can be a general function of this set of proteins.

While current knowledge about RBPs relates to their functions at 5' or 3'-UTRs, it is intriguing that the majority of the conserved RNA folds are found in the CDS of mRNAs. Widespread selection for CDS structural features in both bacteria and eukaryotes has been reported, but the effect of RNA structure remains unknown (Gu et al., 2014). While secondary structures have been

linked to translation inhibition, a recent study evidences that a stem-loop formation within the coding sequence of the *fepA* gene can also activate translation in *E. coli* (Jagodnik, Chiaruttini, & Guillier, 2017). Translation initiation is independent of the nucleotide sequence, but the distance between the start codon and the stem-sloop is decisive for translation initiation rate control (Borujeni et al., 2017). RBPs binding to stem-loop conformations within the N-terminal CDS could possibly contribute to the inhibitory effect of hairpins for example by obstructing 30S binding, inhibiting hairpin unwinding or preventing binding of other important initiation factors. We report a significant number of proteins binding to RNA folds in the coding regions of mRNAs. We not only detect proteins unrelated to RNA-binding such as metabolic enzymes and kinases, but also some classical RBPs such as Pab1 showing surprising binding profiles. Additionally, available PAR-CLIP data and RNA immunoprecipitation experiments validate binding of Pab1 to two RNA folds within the coding sequence of *YEF3* and *RBG2*. Although this is still challenging to understand, our data can help generate new hypotheses that will be important to address in future studies.

Moreover, we examined functional connections between the identified RBPs and mRNA folds using a global yeast genetic interaction dataset (Costanzo et al., 2016). Computational studies established that genetic interactions are indicative of functional relationships among genes as evaluated by GO attributes (Hin et al., 2004). Collective evidences from genetic interaction data and our interactomics screen hint towards functional and mechanistic links of our RBD-mRNA pairs that will be important to evaluate experimentally.

Additionally, we characterized a subset of mRNA-RBP pairs and thus connect structural RNA features to functionality. Using an *in vivo* reporter screen, we explored a possible function in gene expression control by RNA folds in the UTR regions. Of the 12 RNA folds tested, we identified two protein binders as putative regulators. While YDR514C is a yet uncharacterized protein, the other binder, Nsr1, harbors two RNA recognition motifs and is involved in rRNA processing. Previous RNA capture studies also report rRNA-processing proteins

as mRNA-binding proteins (Mitchell et al., 2013; Beckmann et al., 2015; Matia-gonzález et al., 2015). As a complementary strategy, we employ pulsed SILAC to evaluate a role of some mRNA-protein pairs in translational regulation. In agreement with the GFP reporter screen data, Nsr1 scores as a translational repressor on its target mRNA *ATP1* provided no changes were observed at the mRNA level. Next, it will be exciting to focus on *in vivo* mutational studies to delineate the RNA-binding motifs required for binding in more detail.

An unprecedented amount of RNA folding data has been produced in the last years (Thiel, Ochsenreiter, Gadekar, Tanzer, & Hofacker, 2018). Important avenues for future research will be mapping proteins binders to individual RNA folds that link RNA structural features to functionality. Towards this end, our study presents a scalable RNA pull-down approach coupled to quantitative proteomics and provides protein interaction data on 186 evolutionary conserved RNA structures. Furthermore, our resource dataset inspires immediate follow-up projects. For example, addressing the role of Nsr1 in gene expression regulation, functional consequences of Pab1 binding to RNA folds in the coding sequence of mRNAs or a possible role of some tRNA synthetases in translation regulation in yeast. As a next step on the study of RNA structure-mediated function, mutation/rescue studies on different stem-loop formations should be performed.

ACKNOWLEDGEMENTS

I want to thank you ☺ for your trust in all the lab projects and external collaborations I took part in. I also want to acknowledge your patience when piecing together the last manuscript, which has been specially challenging from the distance. I am also very grateful to ☺ for your help throughout this time and the last experiments you performed when I was not able to pipet in Mainz. Also, a big thank you to ☺, who also contributed to the last article during revision.

Thanks to my TAC committee members, ☺ and ☺ for your input during the TAC meetings. I also want to thank my evaluation committee for taking the time to read and evaluate my thesis.

Specially, I am very grateful to you ☺ for your ideas and experimental efforts into the RNA project. Also, I really appreciate your support and discussion during long coffee meetings in the Science Lounge.

A big big thank goes to you, ☺. Thanks for your big smile and for always being so supportive. I probably cannot put down in words how much we shared during our PhD time. All stories come to an end, but I am extremely happy that all our adventures turned into a long-lasting friendship. Cheers to the new adventures to come!

I am also very happy to have met all the lab crew that always brought in a fantastic atmosphere. ☺, ☺, ☺, ☺, ☺ and ☺ thanks for all the over-lunch conversations; sharing frustrations and bringing new ideas; our coffee talks at the terrace, lab music, dancing, zumba, Funzelfahrt… and so many more things we shared during our PhD time. I hope we can continue planning trips together that are always fun! To the German-speaking part of the lab: thanks for your patience in answering my endless grammar questions! ☺, thanks for your support and conversations during our breaks in the terrace. To ☺, ☺ and ☺, I

am grateful you always helped when needed and were very patient with R questions. You guys always organized fun lab day-off activities, introduced Saumagen and russischer Zupfkuchen to us, hosted great barbecues in Dienheim and delicious Christmas cookie baking!

I also want to thank the people I collaborated with at the IMB: ☺, ☺, ☺ and ☺. Also thanks to the support of the **IMB Media Lab** with buffers and media and ☺ from the Flow Cytometry Core Facility for helping out with the reporter screen measurements.

Thanks to the *borrel* team, the PhD student community and coordinators for creating a friendly atmosphere. Also, to all the friendly faces at the IMB that made coffee breaks in the SL very entertaining and my time at the IMB a great experience.

The biggest thank you to you, ☺, for your endless support throughout this time, for creating plots overnight when needed and for countless hours of babysitting.

I also want to express my gratitude to all the people who I forget to mention here but that directly or indirectly contributed to this work.

Moltes gràcies també a la meva **família**: el meu pare, la meva mare, l'☺ i els meus avis, per haver-me ensenyat el valor de l'esforç i resiliència, pel seu suport i paciència durant aquests anys i per totes les hores de cangur que m'han permès tancar aquesta etapa. També als meus amics doctors i no doctors, per no deixar mai de preguntar-me quan acabaria la tesi. I especialment a l'☺, per aportar un toc artístic a la meva portada de tesi.

REFERENCES

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., … Craig Venter, J. (2000). The Genome Sequence of Drosophila melanogaster. *Science*, *287(5461):2185–95*.

Aebersold, R., & Goodlett, D. R. (2001). Mass spectrometry in proteomics. *Chemical Reviews*, *101*(2), 269–295.

Aebersold, Ruedi, & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, *422*.

Aebersold, Ruedi, & Mann, M. (2016, September 14). Mass-spectrometric exploration of proteome structure and function. *Nature*. Nature Publishing Group.

Ahmad, Y., & Lamond, A. I. (2014). A perspective on proteomics in cell biology. *Trends in Cell Biology*. Elsevier Ltd.

Andrews, S., Snowflack, D. R., Clark, I. E., & Gavis, E. R. (2011). Multiple mechanisms collaborate to repress nanos translation in the Drosophila ovary and embryo. *RNA*, *17*(5), 967–977.

Baltz, A. G., Munschauer, M., Schwanhusser, B., Vasile, A., Murakawa, Y., Schueler, M., … Landthaler, M. (2012). The mRNA-Bound Proteome and Its Global Occupancy Profile on Protein-Coding Transcripts. *Molecular Cell*, *46*(5), 674–690.

Bantscheff, M., Lemeer, S., Savitski, M. M., & Kuster, B. (2012, September). Quantitative mass spectrometry in proteomics: Critical review update from 2007 to the present. *Analytical and Bioanalytical Chemistry*.

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.

Baralle, F. E., & Giudice, J. (2017, July 1). Alternative splicing as a regulator of development and tissue identity. *Nature Reviews Molecular Cell Biology*. Nature Publishing Group.

Bartys, N., Kierzek, R., & Lisowiec-Wachnicka, J. (2019). The regulation properties of RNA secondary structure in alternative splicing. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*. Elsevier B.V..

Beati, H., Langlands, A., Have, ten, & Müller, H.-A. J. (2019). SILAC-based quantitative proteomic analysis of Drosophila gastrula stage embryos mutant for fibroblast growth factor signaling. *Fly (Austin)*, *24;1–19*.

Becker, K., Bluhm, A., Casas-Vila, N., Dinges, N., Dejung, M., Sayols, S., … Legewie, S. (2018). Quantifying post-transcriptional regulation in the development of Drosophila melanogaster. *Nature Communications*, *9*(1) doi:10.1038/s41467-018-07455-9.

Beckmann, B. M. (2017). RNA interactome capture in yeast. *Methods*, *118–119*, 82–92.

Beckmann, B. M., Castello, A., & Medenbach, J. (2016). The expanding universe of ribonucleoproteins : of novel RNA-binding proteins and unconventional interactions. *Pflügers Archiv - European Journal of Physiology*, 1029–1040.

Beckmann, B. M., Horos, R., Fischer, B., Castello, A., Eichelbaum, K., Alleaume, A.-M., … Hentze, M. W. (2015a). The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. *Nature Communications*, *6*, 10127.

Beckmann, B. M., Horos, R., Fischer, B., Castello, A., Eichelbaum, K., Alleaume, A. M., … Hentze, M. W. (2015b). uiv. *Nature Communications*, *6* doi:10.1038/ncomms10127.

Bellei, E., Bergamini, S., Monari, E., Fantoni, L. I., Cuoghi, A., Ozben, T., & Tomasi, A. (2011). High-abundance proteins depletion for serum proteomic analysis: Concomitant removal of non-targeted proteins. *Amino Acids*, *40*(1), 145–156.

Benoit, B., He, C. H., Zhang, F., Votruba, S. M., Tadros, W., Westwood, J. T., … Theurkauf, W. E. (2009). An essential role for the RNA-binding protein Smaug during the Drosophila maternal-to-zygotic transition. *Development*, *136*(6), 923–932.

Beynon, R. J., Doherty, M. K., Pratt, J. M., & Gaskell, S. J. (2005). Multiplexed absolute quantification in proteomics using artificial QCAT proteins of concatenated signature peptides. *Nature Methods*, *2*(8), 587–589.

Bier, E., Harrison, M. M., O'connor-Giles, K. M., & Wildonger, J. (2018). Advances in engineering the fly genome with the CRISPR-Cas system. *Genetics*, *208*(1), 1–18.

Boisvert, F. M., Ahmad, Y., Gierliński, M., Charrière, F., Lamont, D., Scott, M., … Lamond, A. I. (2012). A quantitative spatial proteomics analysis of proteome turnover in human cells. *Molecular and Cellular Proteomics*, *11*(3).

Bonaldi, T., Straub, T., Cox, J., Kumar, C., Becker, P. B., & Mann, M. (2008). Combined use of RNAi and quantitative proteomics to study gene function in Drosophila. *Molecular Cell*, *31*(5), 762–772.

Borujeni, A. E., Cetnar, D., Farasat, I., Smith, A., Lundgren, N., & Salis, H. M. (2017). Precise quantification of translation inhibition by mRNA structures that overlap with the ribosomal footprint in N-terminal coding sequences. *Nucleic Acids Research*, *45*(9), 5437–5448.

Brown, J. B., Boley, N., Eisman, R., May, G. E., Stoiber, M. H., Duff, M. O., … Celniker, S. E. (2014). Diversity and dynamics of the Drosophila transcriptome. *Nature*, *512*(7515), 393–399.

Brown, R. S. (2005). Zinc finger proteins: Getting a grip on RNA. *Current Opinion in Structural Biology*. Elsevier Ltd.

Brunner, E., Ahrens, C. H., Mohanty, S., Baetschmann, H., Loevenich, S., Potthast, F., … Aebersold, R. (2007). A high-quality catalog of the Drosophila melanogaster proteome. *Nature Biotechnology*, *25*(5), 576–583.

Butter, F., Scheibe, M., Mö Rl, M., & Mann, M. (2009). *Unbiased RNA-protein interaction screen by quantitative proteomics* (Vol. 106). PNAS.

Calabretta, S., & Richard, S. (2015). Emerging Roles of Disordered Sequences in RNA-Binding Proteins. *Trends in Biochemical Sciences* doi:10.1016/j.tibs.2015.08.012.

Carter, A. P., Clemons, J., Brodersen, D. E., Morgan-Warren, R. J., Hartsch, T., Wimberly, B. T., & Ramakrishnan, V. (2001). Crystal structure of an initiation factor bound to the 30S ribosomal subunit. *Science*, *291*(5503), 498–501.

Casey, J. L., Hentze, M. W., Koeller, D. M., Caughman, S. W., Rouault, T. A., Klausner, R. D., & Harford, J. B. (1988). Iron-responsive elements: regulatory RNA sequences that control mRNA levels and translation. *Science*, *240*(4854), 924–928.

Castello, A., Fischer, B., Eichelbaum, K., Horos, R., Beckmann, B. M., Strein, C., … Hentze, M. W. (2012). Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins. *Cell*, *149*(6), 1393–1406.

Castello, A., Hentze, M. W., & Preiss, T. (2015). Metabolic Enzymes Enjoying New Partnerships as RNA-Binding Proteins. *Trends in Endocrinology and Metabolism*, *26*(12), 746–757.

Catherman, A. D., Skinner, O. S., & Kelleher, N. L. (2014, March 21). Top Down proteomics: Facts and perspectives. *Biochemical and Biophysical Research Communications*. Academic Press Inc.

Chang, Y. C., Tang, H. W., Liang, S. Y., Pu, T. H., Meng, T. C., Khoo, K. H., & Chen, G. C. (2013). Evaluation of Drosophila metabolic labeling strategies for in vivo quantitative proteomic analyses with applications to early pupa formation and amino acid starvation. *Journal of Proteome Research*, *12*(5), 2138–2150.

Chekulaeva, M., & Filipowicz, W. (2009, June). Mechanisms of miRNA-mediated post-transcriptional regulation in animal cells. *Current Opinion in Cell Biology*.

Chintapalli, V. R., Wang, J., & Dow, J. A. T. (2007, June). Using FlyAtlas to identify better Drosophila melanogaster models of human disease. *Nature Genetics*.

Choudhary, C., Kumar, C., Gnad, F., Nielsen, M. L., Rehman, M., Walther, T. C., … Mann, M. (2009). Lysine Acetylation Targets Protein Complexes and Co-Regulates Major Cellular Functions. *Science*, *Vol. 325*(Issue 5942), 834–840.

Chu, C., Zhang, Q. C., Da Rocha, S. T., Flynn, R. A., Bharadwaj, M., Calabrese, J. M., … Chang, H. Y. (2015). Systematic discovery of Xist RNA binding proteins. *Cell*, *161*(2), 404–416.

Cohen-Chalamish, S., Hasson, A., Weinberg, D., Namer, L. S., Banai, Y., Osman, F., & Kaempfer, R. (2009). Dynamic refolding of IFN-γ mRNA enables it to function as PKR activator and translation template. *Nature Chemical Biology*, *5*(12), 896–903.

Conrad, T., Albrecht, A. S., De Melo Costa, V. R., Sauer, S., Meierhofer, D., & Ørom, U. A. (2016). Serial interactome capture of the human cell nucleus. *Nature Communications*, *7*.

Conrads, T. P., Alving, K., Veenstra, T. D., Belov, M. E., Anderson, G. A., Anderson, D. J., … Smith, R. D. (2001). Quantitative analysis of bacterial and mammalian proteomes using a combination of cysteine affinity tags and 15N-metabolic labeling. *Analytical Chemistry*, *73*(9), 2132–2139.

Corbett, A. H. (2018, June 1). Post-transcriptional regulation of gene expression and human disease. *Current Opinion in Cell Biology*. Elsevier Ltd.

Cordero, P., Kladwang, W., Vanlang, C. C., & Das, R. (2012). Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference. *Biochemistry*, *51*(36), 7037–7039.

Costanzo, M., Vandersluis, B., Koch, E. N., Baryshnikova, A., Pons, C., Tan, G., … Boone, C. (2016). A global genetic interaction network maps a wiring diagram of cellular function. *Science*, *353(6306):aaf1420*.

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.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., & Mann, M. (2011a). Andromeda: A peptide search engine integrated into the MaxQuant environment. *Journal of Proteome Research*, *10*(4), 1794–1805.

Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V., & Mann, M. (2011b). Andromeda: A peptide search engine integrated into the MaxQuant environment. *Journal of Proteome Research*, *10*(4), 1794–1805.

Cruz, J. A., & Westhof, E. (2009, February 20). The Dynamic Landscapes of RNA Architecture. *Cell*.

Dassi, E. (2017, September 29). Handshakes and fights: The regulatory interplay of RNA-binding proteins. *Frontiers in Molecular Biosciences*. Frontiers Media S.A..

De Godoy, L. M. F., Olsen, J. V., Cox, J., Nielsen, M. L., Hubner, N. C., Fröhlich, F., … Mann, M. (2008). Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature*, *455*(7217), 1251–1254.

de Godoy, L. M. F., Olsen, J. V., de Souza, G. A., Li, G., Mortensen, P., & Mann, M. (2006). Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. *Genome Biology*, *7*(6).

DeRenzo, C., & Seydoux, G. (2004). A clean start: Degradation of maternal proteins at the oocyte-to-embryo transition. *Trends in Cell Biology*, *14*(8), 420–426.

Dietzl, G., Chen, D., Schnorrer, F., Su, K. C., Barinova, Y., Fellner, M., … Dickson, B. J. (2007). A genome-wide transgenic RNAi library for conditional gene inactivation in Drosophila. *Nature*, *448*(7150), 151–156.

Ding, Y., Tang, Y., Kwok, C. K., Zhang, Y., Bevilacqua, P. C., & Assmann, S. M. (2014). In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. *Nature*, *505*(7485), 696–700.

Dinges, N., Morin, V., Kreim, N., Southall, T. D., & Roignant, J. Y. (2017). Comprehensive Characterization of the Complex lola Locus Reveals a Novel Role in the Octopaminergic Pathway via Tyramine Beta-Hydroxylase Regulation. *Cell Reports*, *21*(10), 2911–2925.

Doerr, A. (2019, January 1). Single-cell proteomics. *Nature Methods*. Nature Publishing Group.

Domon, B., & Aebersold, R. (2006). Mass Spectrometry and Protein Analysis. *Science*, *312(5771):212–7*.

Donnelly, D. P., Rawlins, C. M., DeHart, C. J., Fornelli, L., Schachner, L. F., Lin, Z., … Agar, J. N. (2019). Best practices and benchmarks for intact protein analysis for top-down mass spectrometry. *Nature Methods*, *16*(7), 587–594.

Dou, M., Clair, G., Tsai, C. F., Xu, K., Chrisler, W. B., Sontag, R. L., … Zhu, Y. (2019). High-Throughput Single Cell Proteomics Enabled by Multiplex Isobaric Labeling in a Nanodroplet Sample Preparation Platform. *Analytical Chemistry*.

Dunham, W. H., Mullin, M., & Gingras, A. C. (2012, May). Affinity-purification coupled to mass spectrometry: Basic principles and strategies. *Proteomics*.

Durbin, K. R., Fornelli, L., Fellers, R. T., Doubleday, P. F., Narita, M., & Kelleher, N. L. (2016). Quantitation and Identification of Thousands of Human Proteoforms below 30 kDa. *Journal of Proteome Research*, *15*(3), 976–982.

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.

Fabre, B., Korona, D., Groen, A., Vowinckel, J., Gatto, L., Deery, M. J., … Lilley, K. S. (2016). Analysis of Drosophila melanogaster proteome dynamics during embryonic development by a combination of label-free proteomics approaches. *Proteomics*, *16*(15–16), 2068–2080.

Fairchild, J. N., Walworth, M. J., Horváth, K., & Guiochon, G. (2010). Correlation between peak capacity and protein sequence coverage in proteomics analysis by liquid chromatography-mass spectrometry/mass spectrometry. *Journal of Chromatography A*, *1217*(29), 4779–4783.

Faoro, C., & Ataide, S. F. (2014, October 16). Ribonomic approaches to study the RNA-binding proteome. *FEBS Letters*. Elsevier.

Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F., & Whitehouse, C. M. (1989). Electrospray Ionization for Mass Spectrometry of Large Biomolecules. *Science*, *Vol. 246*(No. 4926), 64–71.

Filipowicz, W., Bhattacharyya, S. N., & Sonenberg, N. (2008, February). Mechanisms of post-transcriptional regulation by microRNAs: Are the answers in sight? *Nature Reviews Genetics*.

Findlay, G. D., Yi, X., MacCoss, M. J., & Swanson, W. J. (2008). Proteomics Reveals Novel Drosophila Seminal Fluid Proteins Transferred at Mating. *Plos Biology*, *6(7):e178*.

Fredens, J., Engholm-Keller, K., Giessing, A., Pultz, D., Larsen, M. R., Højrup, P., … Førgeman, N. J. (2011). Quantitative proteomics by amino acid labeling in C. elegans. *Nature Methods*, *8*(10), 845–847.

Ganser, L. R., Kelly, M. L., Herschlag, D., & Al-Hashimi, H. M. (2019, August 1). The roles of structural dynamics in the cellular functions of RNAs. *Nature Reviews Molecular Cell Biology*. Nature Publishing Group.

Gaspar, I. (2018). *RNA Detection Methods and Protocols Methods in Molecular Biology*.

Geiger, T., Wisniewski, J. R., Cox, J., Zanivan, S., Kruger, M., Ishihama, Y., & Mann, M. (2011). Use of stable isotope labeling by amino acids in cell culture as a spike-in standard in quantitative proteomics. *Nature Protocols*, *6*(2), 147–157.

Gerald M. Rubin, Mark D. Yandell, Jennifer R. Wortman, George L. Gabor, M., Catherine R. Nelson, Iswar K. Hariharan, … Suzanna Lewis. (2000). Comparative Genomics of the Eukaryotes. *Science*, *287(5461):2204–15*.

Gerber, A. P., Herschlag, D., & Brown, P. O. (2004). Extensive association of functionally and cytotopically related mRNAs with Puf family RNA-binding proteins in yeast. *PLoS Biology*.

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. *PNAS*, *100 (12) 6940-6945*.

Gerstberger, S., Hafner, M., & Tuschl, T. (2014). A census of human RNA-binding proteins. *Nature Reviews Genetics*, *15*(12), 829–845.

Gibilisco, L., Zhou, Q., Mahajan, S., & Bachtrog, D. (2016). Alternative Splicing within and between Drosophila Species, Sexes, Tissues, and Developmental Stages. *PLoS Genetics*, *12*(12).

Gilbert, S. (2000). *Developmental Biology* (6th Edition). Sunderland (MA): Sinauer Associates.

Gingras, A. C., Gstaiger, M., Raught, B., & Aebersold, R. (2007, August). Analysis of protein complexes using mass spectrometry. *Nature Reviews Molecular Cell Biology*.

Glisovic, T., Bachorik, J. L., Yong, J., & Dreyfuss, G. (2008). RNA-binding proteins and post-transcriptional gene regulation. *FEBS Letters*, *582*(14), 1977–1986.

Gouw, J. W., Pinkse, M. W. H., Vos, H. R., Moshkin, Y., Verrijzer, C. P., Heck, A. J. R., & Krijgsveld, J. (2009). In vivo stable isotope labeling of fruit flies reveals post-transcriptional regulation in the maternal-to-zygotic transition. *Molecular and Cellular Proteomics*, *8*(7), 1566–1578.

Graveley, B. R., Brooks, A. N., Carlson, J. W., Duff, M. O., Landolin, J. M., Yang, L., … Celniker, S. E. (2011). The developmental transcriptome of Drosophila melanogaster. *Nature*, *471*(7339), 473–479.

Griffin, T. J., Gygi, S. P., Ideker, T., Rist, B., Eng, J., Hood, L., & Aebersold, R. (2002). Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae. *Molecular & Cellular Proteomics : MCP*, *1*(4), 323–333.

Grün, D., Kirchner, M., Thierfelder, N., Stoeckius, M., Selbach, M., & Rajewsky, N. (2014). Conservation of mRNA and protein expression during development of C.elegans. *Cell Reports*, *6*(3), 565–577.

Gu, W., Li, M., Xu, Y., Wang, T., Ko, J. H., & Zhou, T. (2014). The impact of RNA structure on coding sequence evolution in both bacteria and eukaryotes. *BMC Evolutionary Biology*, *14*(1).

Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N., & Altman, S. (1963). *The RNA Moiety of Ribonuclease P Is the Catalytic Subunit of the Enzyme*. *Cell* (Vol. 35).

Gygi, S. P., Beate, R., Gerber, S. A., Turecek, F., Gelb, M. H., & Aebersold, R. (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology*, *17*.

Habchi, J., Tompa, P., Longhi, S., & Uversky, V. N. (2014, July 9). Introducing protein intrinsic disorder. *Chemical Reviews*. American Chemical Society.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., … Tuschl, T. (2010). Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell*, *141*(1), 129–141.

Hall, T. M. T. (2005). Multiple modes of RNA recognition by zinc finger proteins. *Current Opinion in Structural Biology*. Elsevier Ltd.

Havliš, J., & Shevchenko, A. (2004). Absolute quantification of proteins in solutions and in polyacrylamide gels by mass spectrometry. *Analytical Chemistry*, *76*(11), 3029–3036.

Hebert, A. S., Merrill, A. E., Bailey, D. J., Still, A. J., Westphall, M. S., Strieter, E. R., … Coon, J. J. (2013). Neutron-encoded mass signatures for multiplexed proteome quantification. *Nature Methods*, *10*(4), 332–334.

Hein, M. Y., Hubner, N. C., Poser, I., Cox, J., Nagaraj, N., Toyoda, Y., … Mann, M. (2015). A Human Interactome in Three Quantitative Dimensions Organized by Stoichiometries and Abundances. *Cell*, *163*(3), 712–723.

Hentze, M. W., Castello, A., Schwarzl, T., & Preiss, T. (2018, May 1). A brave new world of RNA-binding proteins. *Nature Reviews Molecular Cell Biology*. Nature Publishing Group.

Hentze, M. W., Muckenthaler, M. U., Galy, B., & Camaschella, C. (2010, July). Two to Tango: Regulation of Mammalian Iron Metabolism. *Cell*.

Hin, A., Tong, Y., Lesage, G., Bader, G. D., Ding, H., Xu, H., … Charles Boone, †. (2004). Global Mapping of the Yeast Genetic Interaction Network. *Science*, *Vol. 303*(Issue 5659), 808–813.

Hsu, J. L., & Chen, S. H. (2016, October 28). Stable isotope dimethyl labelling for quantitative proteomics and beyond. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*. Royal Society of London.

Hubner, N. C., Bird, A. W., Cox, J., Splettstoesser, B., Bandilla, P., Poser, I., … Mann, M. (2010). Quantitative proteomics combined with BAC TransgeneOmics reveals in vivo protein interactions. *Journal of Cell Biology*, *189*(4), 739–754.

Hughes, C. S., Moggridge, S., Müller, T., Sorensen, P. H., Morin, G. B., & Krijgsveld, J. (2019). Single-pot, solid-phase-enhanced sample preparation for proteomics experiments. *Nature Protocols*, *14*(1), 68–85.

Huo, X., Ng, J., Tan, M., & Tucker-Kellogg, G. (2019). Genome-Wide Probing of RNA Structure. In *Encyclopedia of Bioinformatics and Computational Biology* (pp. 574–585). Elsevier.

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R. S., & Weissman, J. S. (2009). Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science*, *Vol. 324*(Issue 5924), 218–223.

Jagodnik, J., Chiaruttini, C., & Guillier, M. (2017). Stem-Loop Structures within mRNA Coding Sequences Activate Translation Initiation and Mediate Control by Small Regulatory RNAs. *Molecular Cell*, *68*(1), 158-170.e3.

Jambor, H., Mueller, S., Bullock, S. L., & Ephrussi, A. (2014). A stem-loop structure directs oskar mRNA to microtubule minus ends. *RNA*, *20*(4), 429–439.

Jambor, H., Surendranath, V., Kalinka, A. T., Mejstrik, P., Saalfeld, S., & Tomancak, P. (2014). Systematic imaging reveals features and changing localization of mRNAs in Drosophila development. *ELIFE*, *4:e05003*.

Järvelin, A. I., Noerenberg, M., Davis, I., & Castello, A. (2016, April 6). The new (dis)order in RNA regulation. *Cell Communication and Signaling*. BioMed Central Ltd..

Kalinka, A. T., Varga, K. M., Gerrard, D. T., Preibisch, S., Corcoran, D. L., Jarrells, J., … Tomancak, P. (2010). Gene expression divergence recapitulates the developmental hourglass model. *Nature*, *468*(7325), 811–816.

Karaiskos, N., Wahle, P., Alles, J., Boltengagen, A., Ayoub, S., Kipar, C., … Zinzen, P. (2017). The Drosophila embryo at single-cell transcriptome resolution. *Science*, *Vol. 358*(Issue 6360), 194–199.

Karas Michael, H. F. (1988). Laser Desorption Ionization of Proteins with Molecular Masses exceeding 10,000 daltons. *Analytical Chemistry*, *60*(20), 2299–2301.

Keene, J. D., Komisarow, J. M., & Friedersdorf, M. B. (2006). RIP-Chip: The isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. *Nature Protocols*, *1*(1), 302–307.

Kim, B., & Kim, V. N. (2019). fCLIP-seq for transcriptomic footprinting of dsRNA-binding proteins: Lessons from DROSHA. *Methods*, *152*, 3–11.

Kim, D. I., Jensen, S. C., Noble, K. A., Kc, B., Roux, K. H., Motamedchaboki, K., & Roux, K. J. (2016). An improved smaller biotin ligase for BioID proximity labeling. *Molecular Biology of the Cell*, *27*(8), 1188–1196.

Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M., & Zavolan, M. (2011). A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nature Methods*, *8*(7), 559–567.

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.

Klein, D. J., Moore, P. B., & Steitz, T. A. (2004). The roles of ribosomal proteins in the structure assembly, and evolution of the large ribosomal subunit. *Journal of Molecular Biology*, *340*(1), 141–177.

Klinkert, B., & Narberhaus, F. (2009, August). Microbial thermosensors. *Cellular and Molecular Life Sciences*.

Kozlova, T., & Thummel, C. S. (2002). *Ecdysteroid receptor activation patterns*.

Kramer, K., Sachsenberg, T., Beckmann, B. M., Qamar, S., Boon, K. L., Hentze, M. W., … Urlaub, H. (2014). Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. *Nature Methods*, *11*(10), 1064–1070.

Kretz, M., Siprashvili, Z., Chu, C., Webster, D. E., Zehnder, A., Qu, K., … Khavari, P. A. (2013). Control of somatic tissue differentiation by the long non-coding RNA TINCR. *Nature*, *493*(7431), 231–235.

Krey, J. F., Wilmarth, P. A., Shin, J.-B., Klimek, J., Sherman, N. E., Jeffery, E. D., … Barr-Gillespie, P. G. (2014). Accurate Label-Free Protein Quantitation with High- and Low-Resolution Mass Spectrometers. *J Proteome Res*, *13(2)*, 1034–1044.

Krijgsveld, J., Ketting, R. F., Mahmoudi, T., Johansen, J., Artal-Sanz, M., Peter Verrijzer, C., … Heck, A. J. R. (2003). *Metabolic labeling of C. elegans and D.*

melanogaster for quantitative proteomics. *NATURE BIOTECHNOLOGY* (Vol. 21).

Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., … Greenblatt, J. F. (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, *440*(7084), 637–643.

Kronja, I., Whitfield, Z. J., Yuan, B., Dzeyk, K., Kirkpatrick, J., Krijgsveld, J., & Orr-Weaver, T. L. (2014). Quantitative proteomics reveals the dynamics of protein changes during Drosophila oocyte maturation and the oocyte-to-embryo transition. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(45), 16023–16028.

Krüger, M., Moser, M., Ussar, S., Thievessen, I., Luber, C. A., Forner, F., … Mann, M. (2008). SILAC Mouse for Quantitative Proteomics Uncovers Kindlin-3 as an Essential Factor for Red Blood Cell Function. *Cell*, *134*(2), 353–364.

Kubota, M., Tran, C., & Spitale, R. C. (2015, December 1). Progress and challenges for chemical probing of RNA structure inside living cells. *Nature Chemical Biology*. Nature Publishing Group.

Kulak, N. A., Pichler, G., Paron, I., Nagaraj, N., & Mann, M. (2014). Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nature Methods*, *11*(3), 319–324.

Kuster, B., Schirle, M., Mallick, P., & Aebersold, R. (2005). Scoring proteomes with proteotypic peptide probes. *Nat Rev Mol Cell Biol*, *6(7):577-83*.

Kwok, C. K., Tang, Y., Assmann, S. M., & Bevilacqua, P. C. (2015, April 1). The RNA structurome: Transcriptome-wide structure probing with next-generation sequencing. *Trends in Biochemical Sciences*. Elsevier Ltd.

Kwon, S. C., Yi, H., Eichelbaum, K., Föhr, S., Fischer, B., You, K. T., … Kim, V. N. (2013). The RNA-binding protein repertoire of embryonic stem cells. *Nature Structural and Molecular Biology*.

Lapointe, C. P., Wilinski, D., Saunders, H. A. J., & Wickens, M. (2015). Protein-RNA networks revealed through covalent RNA marks. *Nature Methods*, *12*(12), 1163–1170.

Lau, E., Cao, Q., Lam, M. P. Y., Wang, J., Ng, D. C. M., Bleakley, B. J., … Ping, P. (2018). Integrated omics dissection of proteome dynamics during cardiac remodeling. *Nature Communications*, *9*(1).

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.

Lécuyer, E., Yoshida, H., Parthasarathy, N., Alm, C., Babak, T., Cerovina, T., … Krause, H. M. (2007). Global Analysis of mRNA Localization Reveals a Prominent Role in Organizing Cellular Architecture and Function. *Cell*, *131*(1), 174–187.

Leontis, N. B., Lescoute, A., & Westhof, E. (2006, June). The building blocks and motifs of RNA architecture. *Current Opinion in Structural Biology*.

Leppek, K., Das, R., & Barna, M. (2018, March 1). Functional 5′ UTR mRNA structures in eukaryotic translation regulation and how to find them. *Nature Reviews Molecular Cell Biology*. Nature Publishing Group.

Leppek, K., & Stoecklin, G. (2014). An optimized streptavidin-binding RNA aptamer for purification of ribonucleoprotein complexes identifies novel ARE-binding proteins. *Nucleic Acids Research*, *42*(2).

Li, Z., Adams, R. M., Chourey, K., Hurst, G. B., Hettich, R. L., & Pan, C. (2012). Systematic comparison of label-free, metabolic labeling, and isobaric chemical labeling for quantitative proteomics on LTQ orbitrap velos. In *Journal of Proteome Research* (Vol. 11, pp. 1582–1590).

Lin, R.-J. (2016). *RNA-Protein Complexes and Interactions Methods and Protocols Methods in Molecular Biology 1421*.

Lindemann, C., Thomanek, N., Hundt, F., Lerari, T., Meyer, H. E., Wolters, D., & Marcus, K. (2017). Strategies in relative and absolute quantitative mass spectrometry based proteomics. *Biological Chemistry*, *398*(5–6) doi:10.1515/hsz-2017-0104.

Liu, C., Ma, Y., Shang, Y., Huo, R., & Li, W. (2018, May 1). Post-translational regulation of the maternal-to-zygotic transition. *Cellular and Molecular Life Sciences*. Birkhauser Verlag AG.

Liu, X., Salokas, K., Tamene, F., Jiu, Y., Weldatsadik, R. G., Öhman, T., & Varjosalo, M. (2018). An AP-MS- and BioID-compatible MAC-tag enables comprehensive mapping of protein interactions and subcellular localizations. *Nature Communications*, *9*(1).

Liu, Y., Beyer, A., & Aebersold, R. (2016, April 21). On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell*. Cell Press.

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).

Lunde, B. M., Moore, C., & Varani, G. (2007, June). RNA-binding proteins: Modular design for efficient function. *Nature Reviews Molecular Cell Biology*.

Macek, B., Mann, M., & Olsen, J. V. (2009). Global and Site-Specific Quantitative Phosphoproteomics: Principles and Applications. *Annual Review of Pharmacology and Toxicology*, *49*(1), 199–221.

Makarov, A. (2000). Electrostatic axially harmonic orbital trapping: A high-performance technique of mass analysis. *Analytical Chemistry*, *72*(6), 1156–1162.

Mann, M., Kulak, N. A., Nagaraj, N., & Cox, J. (2013, February 21). The Coming Age of Complete, Accurate, and Ubiquitous Proteomes. *Molecular Cell*.

Marchese, D., de Groot, N. S., Lorenzo Gotor, N., Livi, C. M., & Tartaglia, G. G. (2016). Advances in the characterization of RNA-binding proteins. *Wiley Interdisciplinary Reviews: RNA*, *7*(6), 793–810.

Marondedze, C., Thomas, L., Serrano, N. L., Lilley, K. S., & Gehring, C. (2016). The RNA-binding protein repertoire of Arabidopsis thaliana. *Scientific Reports*, *6*.

Mathews, D. H. (2004). Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, *10*(8), 1178–1190.

Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., & Turner, D. H. (2004). *Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure*. *PNAS* (Vol. 101).

Mathieson, T., Franken, H., Kosinski, J., Kurzawa, N., Zinn, N., Sweetman, G., … Savitski, M. M. (2018). Systematic analysis of protein turnover in primary cells. *Nature Communications*, *9*(1).

Matia-González, A. M., Iadevaia, V., & Gerber, A. P. (2017). A versatile tandem RNA isolation procedure to capture in vivo formed mRNA-protein complexes. *Methods*, *118–119*, 93–100.

Matia-gonzález, A. M., Laing, E. E., & Gerber, A. P. (2015). Conserved mRNA-binding proteomes in eukaryotic organisms. *Nature Publishing Group*, *22*(12), 1027–1033.

Mayr, C. (2017). Regulation by 3-Untranslated Regions doi:10.1146/annurev-genet-120116.

McAlister, G. C., Huttlin, E. L., Haas, W., Ting, L., Jedrychowski, M. P., Rogers, J. C., … Gygi, S. P. (2012). Increasing the multiplexing capacity of TMTs using reporter ion isotopologues with isobaric masses. *Analytical Chemistry*, *84*(17), 7469–7478.

McMahon, A. C., Rahman, R., Jin, H., Shen, J. L., Fieldsend, A., Luo, W., & Rosbash, M. (2016). TRIBE: Hijacking an RNA-Editing Enzyme to Identify Cell-Specific Targets of RNA-Binding Proteins. *Cell*, *165*(3), 742–753.

McShane, E., Sin, C., Zauber, H., Wells, J. N., Donnelly, N., Wang, X., … Selbach, M. (2016). Kinetic Analysis of Protein Stability Reveals Age-Dependent Degradation. *Cell*, *167*(3), 803-815.e21.

Merino, E. J., Wilkinson, K. A., Coughlan, J. L., & Weeks, K. M. (2005). RNA structure analysis at single nucleotide resolution by Selective 2′-Hydroxyl Acylation and Primer Extension (SHAPE). *Journal of the American Chemical Society*, *127*(12), 4223–4231.

Michalski, A., Damoc, E., Hauschild, J.-P., Lange, O., Wieghaus, A., Makarov, A., … Horning, S. (2011). *Q Exactive, a benchtop quadrupole Orbitrap mass*

spectrometer Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer Running title: Q Exactive, a benchtop quadrupole Orbitrap mass analyzer. MCP Papers in Press.

Miillner, E. W., & Kiihn, L. C. (1988). *A Stem-Loop in the 3' Untranslated Region Mediates Iron-Dependent Regulation of Transferrin Receptor mRNA Stability in the Cytoplasm*. *Cell* (Vol. 53).

Mitchell, S. A., Spriggs, K. A., Coldwell, M. J., Jackson, R. J., & Willis, A. E. (2003). *The Apaf-1 Internal Ribosome Entry Segment Attains the Correct Structural Conformation for Function via Interactions with PTB and unr*. *Molecular Cell* (Vol. 11).

Mitchell, S. F., Jain, S., She, M., & Parker, R. (2012). Global analysis of yeast mRNPs. *Nature Structural & Molecular Biology*, *20*(1), 127–133.

Mitchell, S. F., Jain, S., She, M., & Parker, R. (2013). Global analysis of yeast mRNPs. *Nat Struct Mol Biol*, *20*(1), 127–133.

Mittler, G., Butter, F., & Mann, M. (2009). A SILAC-based DNA protein interaction screen that identifies candidate binding proteins to functional DNA elements. *Genome Research*, *19*(2), 284–293.

Müller, T., Kalxdorf, M., Longuespée, R., Kazdal, D. N., Stenzinger, A., & Krijgsveld, J. (2020). Automated sample preparation with SP 3 for low-input clinical proteomics . *Molecular Systems Biology*, *16*(1).

Nagarkar-Jaiswal, S., Lee, P.-T., Campbell, M. E., Chen, K., Anguiano-Zarate, S., Cantu Gutierrez, M., … Duncan, D. (2015). A library of MiMICs allows tagging of genes and reversible, spatial and temporal knockdown of proteins in Drosophila. *ELIFE*, *4:e05338*.

Nahvi, A., Sudarsan, N., Ebert, M. S., Zou, X., Brown, K. L., & Breaker, R. R. (2002). *Genetic Control by a Metabolite Binding mRNA it has become evident that RNA has a significant poten*. *Chemistry & Biology* (Vol. 9).

Nguyen, Q. H., Pervolarakis, N., Nee, K., & Kessenbrock, K. (2018, September 4). Experimental considerations for single-cell RNA sequencing approaches. *Frontiers in Cell and Developmental Biology*. Frontiers Media S.A.

Oeffinger, M. (2012, May). Two steps forward-one step back: Advances in affinity purification mass spectrometry of macromolecular complexes. *Proteomics*.

Olsen, J. V., de Godoy, L. M. F., Li, G., Macek, B., Mortensen, P., Pesch, R., … Mann, M. (2005). Parts per million mass accuracy on an orbitrap mass spectrometer via lock mass injection into a C-trap. *Molecular and Cellular Proteomics*, *4*(12), 2010–2021.

Olsen, J. V., & Mann, M. (2013, December). Status of large-scale analysis of posttranslational modifications by mass spectrometry. *Molecular and Cellular Proteomics*.

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 : MCP*, *1*(5), 376–386.

Ong, S. E., & Mann, M. (2005). Mass Spectrometry–Based Proteomics Turns Quantitative. *Nature Chemical Biology*, *1*(5), 252–262.

Ong, S. E., Mittler, G., & Mann, M. (2004). Identifying and quantifying in vivo methylation sites by heavy methyl SILAC. *Nature Methods*, *1*(2), 119–126.

Oriel, C., & Lasko, P. (2018, July 13). Recent developments in using Drosophila as a model for human genetic disease. *International Journal of Molecular Sciences*. MDPI AG.

Overmyer, K. A., Tyanova, S., Hebert, A. S., Westphall, M. S., Cox, J., & Coon, J. J. (2018). Multiplexed proteome analysis with neutron-encoded stable isotope labeling in cells and mice. *Nature Protocols*, *13*(1), 293–306.

Pappireddi, N., Martin, L., & Wühr, M. (2019, May 15). A Review on Quantitative Multiplexed Proteomics. *ChemBioChem*. Wiley-VCH Verlag.

Peattie, D. A., & Gilbert, W. (1980). *Chemical probes for higher-order structure in RNA (dimethyl sulfate/diethyl pyrocarbonate/base stacking/base pairing/ionic shielding)*. *Biochemistry* (Vol. 77).

Perez-Perri, J. I., Rogell, B., Schwarzl, T., Stein, F., Zhou, Y., Rettel, M., … Hentze, M. W. (2018). Discovery of RNA-binding proteins and characterization of

their dynamic responses by enhanced RNA interactome capture. *Nature Communications*, *9*(1).

Peshkin, L., Wühr, M., Pearl, E., Haas, W., Freeman, R. M., Gerhart, J. C., … Kirschner, M. W. (2015). On the Relationship of Protein and mRNA Dynamics in Vertebrate Embryonic Development. *Developmental Cell*, *35*(3), 383–394.

Pesole, G., Mignone, F., Gissi, C., Grillo, G., Licciulli, F., & Liuni, S. (2001). Structural and functional features of eukaryotic mRNA untranslated regions. *Geme*, *Volume 276*(Issues 1–2), Pages 73-81.

Phan, A. T., Kuryavyi, V., Darnell, J. C., Serganov, A., Majumdar, A., Ilin, S., … Patel, D. J. (2011). Structure-function studies of FMRP RGG peptide recognition of an RNA duplex-quadruplex junction. *Nature Structural and Molecular Biology*, *18*(7), 796–804.

Pierce, A., Unwin, R. D., Evans, C. A., Griffiths, S., Carney, L., Zhang, L., … Whetton, A. D. (2008). Eight-channel iTRAQ enables comparison of the activity of six leukemogenic tyrosine kinases. *Molecular and Cellular Proteomics*, *7*(5), 853–863.

Ramanathan, M., Majzoub, K., Rao, D. S., Neela, P. H., Zarnegar, B. J., Mondal, S., … Khavari, P. A. (2018). RN A-protein interaction detection in living cells. *Nature Methods*, *15*(3), 207–212.

Ramanathan, M., Porter, D. F., & Khavari, P. A. (2019, March 1). Methods to study RNA–protein interactions. *Nature Methods*. Nature Publishing Group.

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.

Ray, D., Kazan, H., Chan, E. T., Castillo, L. P., Chaudhry, S., Talukder, S., … Hughes, T. R. (2009). Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nature Biotechnology*, *27*(7), 667–670.

Rgen Cox, J., Hein, M. Y., Luber, 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* S

Technological Innovation and Resources. *Molecular & Cellular Proteomics*, *13*, 2513–2526.

Ricci, E. P., Kucukural, A., Cenik, C., Mercier, B. C., Singh, G., Heyer, E. E., … Moore, M. J. (2014). Staufen1 senses overall transcript secondary structure to regulate translation. *Nature Structural and Molecular Biology*, *21*(1), 26–35.

Rinn, J. L., & Chang, H. Y. (2012). Genome Regulation by Long Noncoding RNAs. *Annual Review of Biochemistry*, *81*(1), 145–166.

Robles, M. S., Cox, J., & Mann, M. (2014). In-Vivo Quantitative Proteomics Reveals a Key Contribution of Post-Transcriptional Mechanisms to the Circadian Regulation of Liver Metabolism. *PLoS Genetics*, *10*(1).

Roignant, J. Y., & Soller, M. (2017, June 1). m6A in mRNA: An Ancient Mechanism for Fine-Tuning Gene Expression. *Trends in Genetics*. Elsevier Ltd.

Rothstein, J. D., Patel, S., Regan, M. R., Haenggeli, C., Huang, Y. H., Bergles, D. E., … Fisher, P. B. (2005). β-Lactam antibiotics offer neuroprotection by increasing glutamate transporter expression. *Nature*, *433*(7021), 73–77.

Rouskin, S., Zubradt, M., Washietl, S., Kellis, M., & Weissman, J. S. (2014). Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo. *Nature*, *505*(7485), 701–5.

Roux, K. J., Kim, D. I., Raida, M., & Burke, B. (2012). A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *Journal of Cell Biology*, *196*(6), 801–810.

Sarov, M., Barz, C., Jambor, H., Hein, M. Y., Schmied, C., Suchold, D., … Schnorrer, F. (2016). A genome-wide resource for the analysis of protein localisation in Drosophila.

Scheibe, M., Arnoult, N., Kappei, D., Buchholz, F., Decottignies, A., Butter, F., & Mann, M. (2013). Quantitative interaction screen of telomeric repeatcontaining RNA reveals novel TERRA regulators. *Genome Research*, *23*(12), 2149–2157.

Scherrer, T., Mittal, N., & Janga, S. C. (2010). A Screen for RNA-Binding Proteins in Yeast Indicates Dual Functions for Many Enzymes, *5*(11).

Scherrer, T., Mittal, N., Janga, S. C., & Gerber, A. P. (2010). A screen for RNA-binding proteins in yeast indicates dual functions for many enzymes. *PLoS ONE*, *5*(11).

Schwanhäusser, B., Gossen, M., Dittmar, G., & Selbach, M. (2009). Global analysis of cellular protein translation by pulsed SILAC. *Proteomics*, *9*(1), 205–209.

Schwanhüusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., … Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature*, *473*(7347), 337–342.

Scigelova, M., & Makarov, A. (2006). Orbitrap mass analyzer - Overview and applications in proteomics. In *Proteomics* (Vol. 1, pp. 16–21).

Shchepachev, V., Bresson, S., Spanos, C., Petfalski, E., Fischer, L., Rappsilber, J., & Tollervey, D. (2019). Defining the RNA interactome by total RNA -associated protein purification . *Molecular Systems Biology*, *15*(4).

Shevchenko, A., Tomas, H., Havliš, J., Olsen, J. V., & Mann, M. (2007). In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nature Protocols*, *1*(6), 2856–2860.

Shukla, A. K., & Futrell, J. H. (2000). *Tandem mass spectrometry: dissociation of ions by collisional activation*. *JOURNAL OF MASS SPECTROMETRY J. Mass Spectrom* (Vol. 35).

Sleno, L., & Volmer, D. A. (2004, October). Ion activation methods for tandem mass spectrometry. *Journal of Mass Spectrometry*.

Srisawat, C., & Engelke, D. R. (2001). *Streptavidin aptamers: Affinity tags for the study of RNAs and ribonucleoproteins*.

Srisawat, C., Goldstein, I. J., & Engelke, D. R. (2001). *Sephadex-binding RNA ligands: rapid affinity purification of RNA from complex RNA mixtures*. *Nucleic Acids Research* (Vol. 29).

St Johnston, D. (2002). The art and design of genetic screens: Drosophila melanogaster. *Nature Reviews Genetics*.

Stadlmeier, M., Bogena, J., Wallner, M., Wühr, M., & Carell, T. (2018). Ein auf Sulfoxid basierendes, isobares Derivatisierungsreagens für die präzise

quantitative Massenspektrometrie. *Angewandte Chemie*, *130*(11), 3008–3013.

Steen, H., & Mann, M. (2004, September). The ABC's (and XYZ's) of peptide sequencing. *Nature Reviews Molecular Cell Biology*.

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., … Smibert, P. (2017). Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, *14*(9), 865–868.

Sugimoto, Y., König, J., Hussain, S., Zupan, B., Curk, T., Frye, M., & Ule, J. (2012). Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions. *Genome Biology*, *13*(8), R67.

Sun, L., Bertke, M. M., Champion, M. M., Zhu, G., Huber, P. W., & Dovichi, N. J. (2015). Quantitative proteomics of Xenopus laevis embryos: Expression kinetics of nearly 4000 proteins during early development. *Scientific Reports*, *4*.

Sury, M. D., Chen, J. X., & Selbach, M. (2010). The SILAC fly allows for accurate protein quantification in vivo. *Molecular and Cellular Proteomics*, *9*(10), 2173–2183.

Susan E. Celniker, Laura A. L. Dillon, Mark B. Gerstein, Kristin C. Gunsalus, Steven Henikoff, Gary H. Karpen, … modENCODE Consortium. (2009). *Unlocking the secrets of the genome*.

Swaminathan, J., Boulgakov, A. A., Hernandez, E. T., Bardo, A. M., Bachman, J. L., Marotta, J., … Marcotte, E. M. (2018). Highly parallel single-molecule identification of proteins in zeptomole-scale mixtures. *Nature Biotechnology*, *36*(11), 1076–1091.

Sysoev, V. O., Fischer, B., Frese, C. K., Gupta, I., Krijgsveld, J., Hentze, M. W., … Ephrussi, A. (2016). Global changes of the RNA-bound proteome during the maternal-to-zygotic transition in Drosophila. *Nature Communications*, *7*.

Tadros, W., & Lipshitz, H. D. (2009). The maternal-to-zygotic transition: A play in two acts. *Development*, *136*(18), 3033–3042.

Thakur, S. S., Geiger, T., Chatterjee, B., Bandilla, P., Fröhlich, F., Cox, J., & Mann, M. (2011). Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. *Molecular and Cellular Proteomics*, *10*(8).

Thandapani, P., O'Connor, T. R., Bailey, T. L., & Richard, S. (2013, June 6). Defining the RGG/RG Motif. *Molecular Cell*.

Thiel, B. C., Ochsenreiter, R., Gadekar, V. P., Tanzer, A., & Hofacker, I. L. (2018). RNA structure elements conserved between mouse and 59 other vertebrates. *Genes*, *9*(8).

Tholey, A., & Becker, A. (2017, November 1). Top-down proteomics for the analysis of proteolytic events - Methods, applications and perspectives. *Biochimica et Biophysica Acta - Molecular Cell Research*. Elsevier B.V.

Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., … 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.

Thul, P. J., Akesson, L., Wiking, M., Mahdessian, D., Geladaki, A., Ait Blal, H., … Lundberg, E. (2017). A subcellular map of the human proteome. *Science*, *356*(6340).

Thurmond, J., Goodman, J. L., Strelets, V. B., Attrill, H., Gramates, L. S., Marygold, S. J., … Baker, P. (2019). FlyBase 2.0: The next generation. *Nucleic Acids Research*, *47*(D1), D759–D765.

Timp, W., & Timp, G. (2020). *Beyond mass spectrometry, the next step in proteomics*.

Toby, T. K., Fornelli, L., & Kelleher, N. L. (2016). Progress in Top-Down Proteomics and the Analysis of Proteoforms. *Annual Review of Analytical Chemistry*, *9*(1), 499–519.

Tomancak, P., Berman, B. P., Beaton, A., Weiszmann, R., Kwan, E., Hartenstein, V., … Rubin, G. M. (2007). Global analysis of patterns of gene expression during Drosophila embryogenesis. *Genome Biology*, *8*(7).

Trendel, J., Schwarzl, T., Horos, R., Prakash, A., Bateman, A., Hentze, M. W., & Krijgsveld, J. (2019). The Human RNA-Binding Proteome and Its Dynamics during Translational Arrest. *Cell*, *176*(1–2), 391-403.e19.

Tress, M. L., Bodenmiller, B., Aebersold, R., & Valencia, A. (2008). Proteomics studies confirm the presence of alternative protein isoforms on a large scale. *Genome Biology*, *9*(11).

Tsai, B. P., Wang, X., Huang, L., & Waterman, M. L. (2011). Quantitative profiling of in vivo-assembled RNA-protein complexes using a novel integrated proteomic approach. *Molecular and Cellular Proteomics*, *10*(4).

Tsvetanova, N. G., Klass, D. M., Salzman, J., & Brown, P. O. (2010). Proteome-wide search reveals unexpected RNA-binding proteins in saccharomyces cerevisiae. *PLoS ONE*, *5*(9), 1–12.

Tyanova, S., Temu, T., & Cox, J. (2016). The MaxQuant computational platform for mass spectrometry – based shotgun proteomics. *Nature Protocols*, *11*(12), 2301–2319.

Urdaneta, E. C., Vieira-Vieira, C. H., Hick, T., Wessels, H. H., Figini, D., Moschall, R., … Beckmann, B. M. (2019). Purification of cross-linked RNA-protein complexes by phenol-toluol extraction. *Nature Communications*, *10*(1).

Van Nostrand, E. L., Gelboin-Burkhart, C., Wang, R., Pratt, G. A., Blue, S. M., & Yeo, G. W. (2017). CRISPR/Cas9-mediated integration enables TAG-eCLIP of endogenously tagged RNA binding proteins. *Methods*, *118–119*, 50–59.

Visscher, M., De Henau, S., Wildschut, M. H. E., van Es, R. M., Dhondt, I., Michels, H., … Dansen, T. B. (2016). Proteome-wide Changes in Protein Turnover Rates in C. elegans Models of Longevity and Age-Related Disease. *Cell Reports*, *16*(11), 3041–3051.

Walther, T. C., & Mann, M. (2010, August 23). Mass spectrometry-based proteomics in cell biology. *Journal of Cell Biology*. Rockefeller University Press.

Wan, Y., Qu, K., Zhang, Q. C., Flynn, R. A., Manor, O., Ouyang, Z., … Chang, H. Y. (2014). Landscape and variation of RNA secondary structure across the human transcriptome. *Nature*, *505*(7485), 706–709.

Wang, X., Mclachlan, J., Zamore, P. D., & Tanaka Hall, T. M. (2002). *Forbes and proteins generally*. *Cell* (Vol. 110). Ahringer and Kimble.

Warf, M. B., & Berglund, J. A. (2010, March). Role of RNA structure in regulating pre-mRNA splicing. *Trends in Biochemical Sciences*.

Washburn, M. P., Wolters, D., & Yates III, J. R. (2001). *Large-scale analysis of the yeast proteome by multidimensional protein identification technology*.

Webb-Robertson, B. J. M., Wiberg, H. K., Matzke, M. M., Brown, J. N., Wang, J., McDermott, J. E., … Waters, K. M. (2015, May 1). Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *Journal of Proteome Research*. American Chemical Society.

Weisser, H., Nahnsen, S., Grossmann, J., Nilse, L., Quandt, A., Brauer, H., … Malmström, L. (2013). An automated pipeline for high-throughput label-free quantitative proteomics. *Journal of Proteome Research*, *12*(4), 1628–1644.

Werner, T., Becher, I., Sweetman, G., Doce, C., Savitski, M. M., & Bantscheff, M. (2012). High-resolution enabled TMT 8-plexing. *Analytical Chemistry*, *84*(16), 7188–7194.

Westhof, E. (2015, April 1). Twenty years of RNA crystallography. *RNA*. Cold Spring Harbor Laboratory Press.

Westman-Brinkmalm, A., Abramsson, A., Pannee, J., Gang, C., Gustavsson, M. K., von Otter, M., … Zetterberg, H. (2011). SILAC zebrafish for quantitative analysis of protein turnover and tissue regeneration. *Journal of Proteomics*, *75*(2), 425–434.

Wiesner, J., Premsler, T., & Sickmann, A. (2008, November). Application of electron transfer dissociation (ETD) for the analysis of posttranslational modifications. *Proteomics*.

Wilk, R., Hu, J., Blotsky, D., & Krause, H. M. (2016). RESOURCE/METHODOLOGY Diverse and pervasive subcellular distributions for both coding and long noncoding RNAs.

Wilkinson, K. A., Merino, E. J., & Weeks, K. M. (2006). Selective 2′-hydroxyl acylation analyzed by primer extension (SHAPE): Quantitative RNA

structure analysis at single nucleotide resolution. *Nature Protocols*, *1*(3), 1610–1616.

Winkler, W., Nahvi All, & Greaker, R. R. (2002). Thiamine derivatives bind messenger RNAs directly to reglate bacterial gene expression. *Nature*, *419*.

Wiśniewski, J. R., Zougman, A., Nagaraj, N., & Mann, M. (2009). Universal sample preparation method for proteome analysis. *Nature Methods*, *6*(5), 359–362.

Wolf-Yadlin, A., Hu, A., & Noble, W. S. (2016). Technical advances in proteomics: New developments in data-independent acquisition. *F1000Research*, *5*.

Wolpert, L., Tickle, C., & Martinez Arias, A. (2015). *PRINCIPLES OF DEVELOPMENT* (FIFTH). OXFORD UNIVERSITY PRESS.

Wühr, M., Haas, W., McAlister, G. C., Peshkin, L., Rad, R., Kirschner, M. W., & Gygi, S. P. (2012). Accurate multiplexed proteomics at the MS2 level using the complement reporter ion cluster. *Analytical Chemistry*, *84*(21), 9214–9221.

Xian, F., Hendrickson, C. L., & Marshall, A. G. (2012). High resolution mass spectrometry. *Analytical Chemistry*, *84*(2), 708–719.

Xing, X., Zhang, C., Li, N., Zhai, L., Zhu, Y., Yang, X., & Xu, P. (2014). Qualitative and quantitative analysis of the adult Drosophila melanogaster proteome. *Proteomics*, *14*(2–3), 286–290.

Zeiler, M., Straube, W. L., Lundberg, E., Uhlen, M., & Mann, M. (2012). A protein epitope signature tag (PrEST) library allows SILAC-based absolute quantification and multiplexed determination of protein copy numbers in cell lines. *Molecular and Cellular Proteomics*, *11*(3).

Zeng, F., Peritz, T., Kannanayakal, T. J., Kilk, K., Eirílksdóttir, E., Langel, U., & Eberwine, J. (2006). A protocol for PAIR: PNA-assisted identification of RNA binding proteins in living cells. *Nature Protocols*, *1*(2), 920–927.

Zhang, Z., Krauchunas, A. R., Huang, S., & Wolfner, M. F. (2018). Maternal proteins that are phosphoregulated upon egg activation include crucial factors for oogenesis, egg activation and embryogenesis in Drosophila melanogaster. *G3: Genes, Genomes, Genetics*, *8*(9), 3005–3018.

Zhao, B. S., Roundtree, I. A., & He, C. (2016, December 19). Post-transcriptional gene regulation by mRNA modifications. *Nature Reviews Molecular Cell Biology*. Nature Publishing Group.

Zubarev, R. A. (2004). Electron-capture dissociation tandem mass spectrometry. *Current Opinion in Biotechnology*. Elsevier Ltd.

LIST OF ABBREVIATIONS

3D	Three-dimensional
AP	Affinity purification
AQUA	Absolute quantification
Arg / R	Arginine
BAC	Bacterial artificial chromosome
BirA	Biotin ligase
C-terminus	Carboxy terminus
C18	Octadecyl carbon chain
CDS	Coding sequence
CHART	Capture hybridization analysis of RNA targets
ChIRP-MS	Chromatin isolation by RNA purification
CITE-seq	Cellular indexing of transcriptomes and epitopes by sequencing
CLIP	Cross-linking immunoprecipitation
DC	Apply constant
DDA	Data-dependent aquisition
DIA	Data-independent aquisition
DML	Dimethyl labeling
DMS	Dimethyl sulfate
DTT	Dithiothreitol
ECD	Electron capture dissociation
ESI	Electrospray ionization
ETD	Electron transfer dissociation
FASP	Filter-Aided Sample Preparation
FDR	False discovery rate
FT-ICR	Fourier transform ion cyclotron resonance
GFP	Green fluorescence protein
h	Hours
HA	Hemagglutinin
HCD	Higher-energy C-trap dissociation
HPLC	High performance liquid chromatography
iBAQ	Intensity-based absolute quantification
ICAT	Isotope-coded affinity tag
IMAC	Ion-metal affinity chromatography
IMS	Ion mobility spectrometry

INFG	Interferon gamma
IRE	Iron responsive element
IRES	Internal ribosomal entry site
IRP1	Iron regulatory protein 1
iTRAQ	Isobaric tags for relative and abosulte quantification
kDa	Kilo Dalton
LC	Liquid chromatography
LFQ	Label-free quantification
Lys / K	Lysine
M	Molar
m^6A	N6-methyladenosine
MALDI	Matrix-assisted laser desorption/ionization
mDa	Milli Dalton
min	Minutes
miRNA	MicroRNA
mRNA	Messenger RNA
mRNP	Messenger ribonucleoprotein particle
MS	Mass spectrometry
N-terminus	Amino terminus
NaBH3CN	Sodium cyanoborohydride
PAGE	Polyacrylamide gel electrophoresis
PAIR	Peptide-nucleic-acid-assisted identification of RBPs
PAS	Polyadenylation site
PEP	Posterior error probability
ppm	Parts per million
PSM	Peptide spectrum match
Ptex	Phenol toluol extraction
PTGR	Post-transcriptional gene regulation
PTM	Post-translational modifications
PUF	Pumilio
RAP	RNA affinity purification
RBP	RNA-binding protein
RF	Alternating voltages
RIP	RNA immunoprecipitation
RRM	RNA recognition motif
SCX	Strong-cation-exchange chromatography

SDS	Sodium dodecyl sulfate
SHAPE	Selective 2'-hydroxyl acylation with primer extension
SILAC	Stable isotope labeling by amino acids in cell culture
SMN1	Survival motor neuron 1
SRM/MRM	Single/multi reaction monitoring
TAP	Tandem affinity purification
TFA	Trifluoracetic acid
TiO2	Titanium dioxide
TMT	Tandem mass tag
TOF	Time of flight
TRIBE	Targets of RNA-binding proteins identified by editing
TRIP	Tandem RNA isolation procedure
uORF	Upstream open reading frame
UTR	Untranslated region
UV	Ultra violet
ZNF	Zinc finger domain