Quantitative Modeling of RNA-Protein Interactions

Dissertation

for the award of the degree
"Doctor rerum naturalium"
of the Georg-August-Universität Göttingen

within the doctoral program
International Max Planck Research School
for Molecular Biology (IMPRS-MolBio)
of the Georg-August University School of Science (GAUSS)

submitted by
Salma Sohrabi-Jahromi
from Jahrom, Iran
Göttingen, 2021

Thesis Advisory Committee

- Dr. Johannes Söding

  Research Group Quantitative and Computational Biology, Max Planck Institute for Biophysical Chemistry

- Prof. Dr. Henning Urlaub

  Research Group Bioanalytical Mass Spectrometry, Max Planck Institute for Biophysical Chemistry

- Prof. Dr. Michael Habeck

  Research Group Statistical Inverse Problems in Biophysics, Max Planck Institute for Biophysical Chemistry (Current affiliation: Microscopic Image Analysis Group, University Hospital Jena)

Members of the Examination Board

1st Reviewer: Dr. Johannes Soeding

Research Group Quantitative and Computational Biology, Max Planck Institute for Biophysical Chemistry

2nd Reviewer: Prof. Dr. Henning Urlaub

Research Group Bioanalytical Mass Spectrometry, Max Planck Institute for Biophysical Chemistry

Further members of the Examination Board

- Prof. Dr. Michael Habeck

  Research Group Statistical Inverse Problems in Biophysics, Max Planck Institute for Biophysical Chemistry (Current affiliation: Microscopic Image Analysis Group, University Hospital Jena)

- Dr. Juliane Liepe

  Research Group Quantitative and Systems Biology, Max Planck Institute for Biophysical Chemistry

- Prof. Dr. Burkhard Morgenstern

  Institute for Microbiology and Genetics, Department Bioinformatics, Georg-August University Göttingen

- Prof. Dr. Argyris Papantonis

  Institute of Pathology, University Medical Center Göttingen

Date of oral examination: April 12, 2021

Acknowledgements

Science is a team sports. While I will narrate this work from my own perspective, it is important to emphasize that throughout my career I have been deeply inspired, moved, and elevated by people around me, without whom this would not have been possible:

First and foremost, I thank Dr. Johannes Söding for believing in me and for giving me the opportunity to learn and grow in his group. I am grateful for the space and freedom he provided me to pursue my own ideas and develop myself as an independent thinker. I appreciate all the nice discussions we had (scientific and beyond), and for sharing his love for science with me. I have learnt a lot from him.

I thank the members of my thesis advisory committee, Prof. Dr. Henning Urlaub and Prof. Dr. Michael Habeck, for their constant support, guidance, and good ideas in the last years. Furthermore, I am thankful to Dr. Juliane Liepe, Prof. Dr. Burkhard Morgenstern, and Prof. Dr. Argyris Papantonis for taking their time and agreeing to be part of my examination board.

I would like to thank Christian Roth, for all the scientific discussions and for his constant support and advice during my doctoral work. He has been instrumental in my endeavors to become a better programmer and has encouraged me to take bigger challenges than those I thought were possible.

My deep appreciation goes to Dr. Steffen Burkhardt for giving me the opportunity to pursue my dreams in Germany, for his constant support in the last six years, for giving me the freedom and encouragement to implement my ideas for our graduate school, and for introducing me to many brilliant minds that continue to inspire me every day.

Much of this thesis reports on results that were obtained in an extremely collaborative and interactive environment. This work would not have been possible without the dedication and curiosity of my collaborators. I am grateful to Prof. Dr. Patrick Cramer for his guidance, determination, and commitment to push the RNA degradation project, as well as Dr. Katharina Hoffman for our scientific discussions. I also had many stimulative and rewarding discussions with Dr. Johannes Söding, Dr. David Zwicker, and Dr. Marc Böhning about biomolecular condensates, and their regulation. These discussions expanded my horizons on how basic physical forces govern many aspects of cellular biology. I am very grateful for being part of the condensate club. A big thanks also goes to our overseas collaborators Prof. Steven Hahn and Dr. Ariel Erijman for the inspiring scientific exchanges and for giving me the opportunity to work on the transcription activation project. This was my first encounter with deep neural networks which helped me appreciate their power and seeded my interest and curiosity.

I have greatly benefited from the wonderful working atmosphere in the Söding group. I thank Ruoshi Zhang for introducing me to the fine art of espresso making and Chinese cooking, Milot

Mirdita for always volunteering his technical support and hosting our game nights and dinners, Dr. Saikat Banerjee for organizing our memorable hikes, Dr. Eli Levy Karin and Dr. Nikolaos Papadopoulos for the nice discussions and their friendship, Annika Jochheim for being a caring office neighbor, Dr. Wanwan Ge, Dr. Franco Simonetti, Dr. Gonzalo Parra, Dr. Clovis Galiez, Dr. Martin Steinegger, Étienne Morice, and all master and bachelor students for the overall kind atmosphere and the great coffee break discussions. I am also thankful to the current and past members of the Cramer group, particularly Dr. Marc Böhning, Dr. Sara Osman, Dr. Saskia Gressel, Dr. Dmitry Tegunov, Dr. Björn Schwalb, and Dr. Michael Lidschreiber for their friendship, support, philosophical discussions, and scientific exchanges.

I am very grateful to Viktoriia Huryn, Simon Stitzinger, Matthew Grieshop, Griorgos Kallergis, and Florian Kriegel for trusting me as their supervisor and giving me the opportunity to grow as a teacher and mentor. I have learned a lot from every one of them and found the greatest satisfaction in seeing them outgrow their challenges. It was a pleasure working with them and I wish them the best of luck for their future career.

I would like to thank Dr. Marc Böhning, Ruoshi Zhang, and Christian Roth for reviewing parts of this document.

I am thankful to Kerstin Grüniger and Dr. Steffen Burkhardt for making the graduate experience smooth and rewarding, as well as all current and past members of the IMPRS-MolBio program for creating an extraordinary and welcoming environment which fostered many deep friendships. I thank Laura Ahumada-Arranz, Valentina Manzini, Rashi Goel, Katarina Harasimov, Kristina Stakyte, Volodymyr Mykhailiuk, Deniz Kaya, and other members of my MolBio class for giving me the company, support and understanding I needed during my studies.

Last but not least, I am extremely grateful to my family for encouraging me to take bigger challenges and for their unconditional support that made this journey possible.

Summary

RNA-binding proteins (RBPs) impact every aspect of RNA metabolism including RNA transcription, maturation, export, localization, translation, and stability. Specific RNA-protein interactions therefore play a central role in regulating many cellular processes. However, most RBPs preferentially bind short, often degenerate sequence motifs (∼3-5 bases) that alone cannot explain how they target only specific subsets of transcripts in the cell. In this thesis, I report on the analysis and the thermodynamic modeling of RNA-protein interaction datasets, with the aim of cracking the code behind RBP specificity.

In the first part of my dissertation, I examine RBPs involved in the general eukaryotic RNA degradation pathway. We generated transcriptome-wide maps of RNA-protein interactions in yeast for 30 yeast RNA decay factors using photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP). In-depth bioinformatic analysis revealed that the decay machineries responsible for degradation of the two RNA ends differ in their substrate specificity. We identified TRAMP4 and exosome as the main complexes involved in Nrd1/Nab3 mediated RNA degradation. Moreover, modeling the dependence of mRNA half-life on degradation factor binding suggested that the recruitment of decapping factors happens only upon RNA degradation, while other decay factors may already associate with mRNAs earlier for their surveillance. Furthermore, global comparison of RNA-binding profiles of decay factors with those of other RNA processing proteins indicated many functional associations with the decay factors.

In the second part of this thesis, I introduce Bipartite Motif Finder (BMF), a computational tool that adopts thermodynamic modeling for the discovery of multivalent RNA-protein interactions. Many RBPs have multiple domains that allow them to target multiple short RNA sequences simultaneously in a cooperative manner, others may achieve cooperativity through oligomerization. This results in specificities and affinities that can be many orders of magnitude higher than those possible by single-domain binding events. Yet, previously available motif discovery approaches have not taken this cooperativity into account. BMF takes full account of the cooperativity and calculates binding probabilities by the weighted sum of all binding configurations determined through thermodynamic modeling. By applying BMF on a high-throughput RNA SELEX (HTR-SELEX) dataset of 78 RBPs, we show that bipartite binding is widespread and that the two motif cores are often similar and low in sequence complexity. We also show that BMF can learn the spatial geometry between the binding sites and predict new RBP binding sites in transcripts with an accuracy competitive with existing motif discovery approaches. We made BMF easily accessible for computationally inexperienced users via the web server (https://bmf.soedinglab.org). BMF source code is also available under a GPL license (https://github.com/soedinglab/bipartite_motif_finder).

Contents

List of commonly used abbreviations

BMF Bipartite Motif Finder

cDNA complementary DNA

CUT Cryptic Unstable Transcripts

HTR-SELEX High-Throughput RNA Systematic Evolution of Ligands by EXponential enrichment

iCLIP individual-nucleotide-resolution Cross-Linking ImmunoPrecipitation

LLPS Liquid-Liquid Phase Separation

mRNA messenger RNA

mRNP messenger RiboNucleoProtein

ncRNA non-coding RNA

NUT Nrd1-Unterminated Transcript

PAR-CLIP PhotoActivatable-Ribonucleoside-Enhanced Cross-Linking ImmunoPrecipitation

Pol II RNA polymerase II

pre-mRNA precursor mRNA

RBD RNA Binding Domain

RBP RNA Binding Protein

rRNA ribosomal RNA

snRNA small nuclear RNA

snoRNA small nucleolar RNA

SUT Stable Unannotated Transcript

TF Transcription Factor

tRNA transfer RNA

UTR UnTranslated region ($3'$ and $5'$ UTRs in mRNA)

1 Introduction

The genetic information that makes up the human body is encoded in about three billion base pairs of deoxyribonucleic acid (DNA) (Dahm, 2005; Venter et al., 2001). Complex biological processes read this genetic information in order to decode various features of our cells, such as the concentrations of their inner molecules and consequently their growth, shape and function (Levine and Tjian, 2003; Hager et al., 2009). The flow of genetic information is described in the *central dogma of molecular biology*: double-stranded DNA is transcribed to single-stranded ribonucleic acid (RNA) molecules which are subsequently translated into proteins (Figure 1.1)(Crick, 1970). In transcription, the first step of the process, RNA polymerases bind control regions at the beginning of genes (transcribed genomic regions) and copy the gene's information into single-stranded RNA molecules (Roeder, 2019; Cramer, 2019). These RNA molecules (transcripts) provide the instructions to produce polypeptide sequences which fold into functional proteins, in a process called translation (Crick, 1958; Ramakrishnan, 2002). Apart from functioning as protein blueprints, RNAs can serve as enzymes, protein scaffolds, or regulators in cellular processes (Mattick and Makunin, 2006; Eddy, 2001; Nam et al., 2016). RNAs are therefore the central macromolecules that bridge the genomic information to cellular function. Hence, it is essential for the cell to control the rate of RNA synthesis, degradation, and RNA localization in response to environment stimuli or in the process of development (Licatalosi and Darnell, 2010; Shyu et al., 2008). In order to facilitate regulation of RNA functions, these molecules are never naked in the cell. Their cellular level, location, and chemical modifications are tightly controlled by RNA-binding proteins (RBPs) that can target RNA molecules specifically and thereby control their fate (Gerstberger et al., 2014; Mitchell and Parker, 2014; Müller-McNicoll and Neugebauer, 2013).

Understanding the complex interplay between RNA molecules and their regulating proteins is the topic of my doctoral research. In this thesis, I will discuss new insights into RNA-protein interactions in the context of RNA degradation. Furthermore, I will introduce computational and thermodynamic approaches for modeling the cooperative nature of RNA-protein interactions. Finally, I will illustrate the contributions our model makes towards understanding specificity in the context of RNA recognition by proteins. In order to provide the reader with the necessary information needed to understand this work, I first review the various stages of RNA metabolism, highlighting RNA degradation in greater depth (section 1.1). Then I will explain mechanisms that allow proteins to target RNA molecules with specificity (section 1.2), and summarize experimental and quantitative approaches for studying RNA-protein interactions (section 1.3). Finally, I will outline the scope of this thesis and enumerate the scientific questions addressed by this work (1.4).

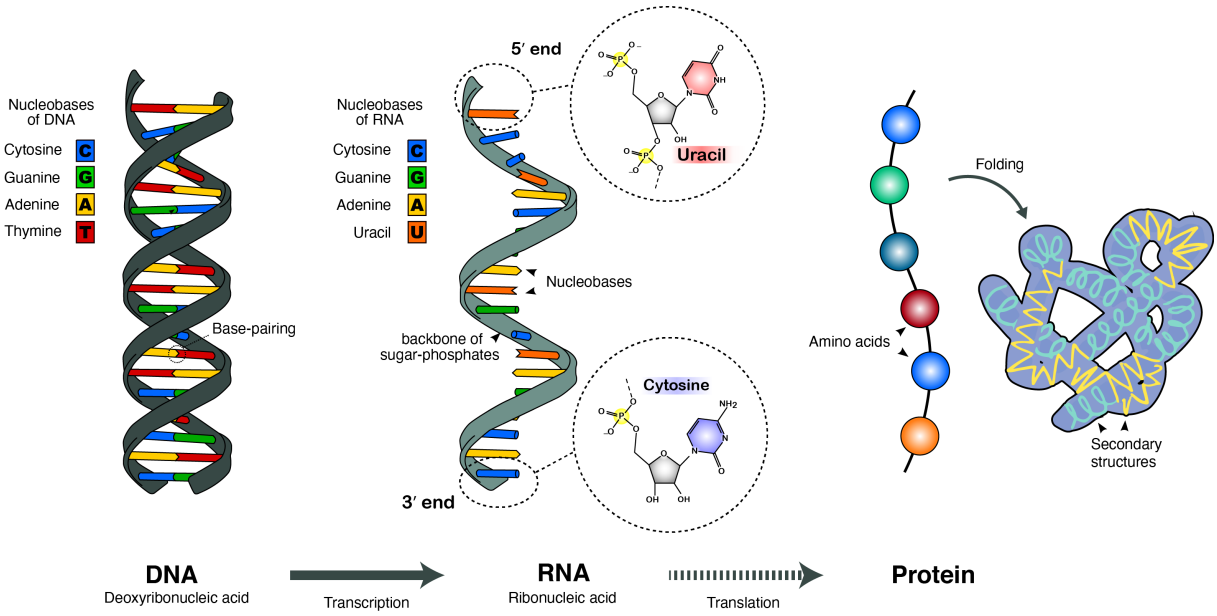


Figure 1.1: **The central dogma of molecular biology.** The central dogma describes the flow of genetic information from DNA to proteins in the cell. The information is stored in the form of double-stranded polynucleotide DNA molecules, encoded as a sequence of four nucleobases: cytosine (C), guanine (G), adenine (A), and thymine (T). During transcription, RNA polymerases create single-stranded RNA molecules using DNA as a template. In this process, the RNA polymerase matches RNA nucleotides to the same base in the DNA sequence with the exception of thymine, which is substituted by uracil (U). A phosphate group attached to the fifth carbon in the sugar-ring marks the beginning of the RNA chain (5′ end), while the end of the RNA molecule is marked with the hydroxyl group of the third carbon in the sugar-ring (3′ end). Some RNA molecules provide the instructions for protein synthesis, a process called translation. During this process a chain of amino acids (also called the polypeptide chain) is produced by matching nucleotide triplets of RNA to their encoding amino acids. The polypeptide then folds into its final structure and can perform its cellular function. (Figure is adapted from Wikipedia)

1.1 The complex life of eukaryotic RNAs

RNA molecules are polymers of four nucleotides, defined by specific nucleobases: cytosine (C), guanine (G), adenine (A), and uracil (U) (Figure 1.1). RNA is not a symmetric polymer but maintains directionality: its first nucleotide contains a phosphate group attached to the fifth carbon in its ribose sugar-ring (hence called the 5′ end), and its last nucleotide is marked with the hydroxyl group of the third carbon in its sugar-ring (hence called the 3′ end) (RajBhandary, 1968). RNA molecules are uniquely versatile as they not only have the ability to store genetic information, but they can also fold into complex three-dimensional structures (Rich and Davies, 1956; Holley et al., 1965; Wan et al., 2011), receive various molecular modifications to modulate their function (Cantara et al., 2010; Boo and Kim, 2020; Kiss, 2001), have enzymatic activity (Lincoln and Joyce, 2009; Haseloff and Gerlach, 1988), and act as a scaffold to recruit molecules needed to build a biological machinery (Zappulla and Cech, 2004; Tsai et al., 2010; Fox et al., 2018a,b). This versatility has made RNA not only a prime candidate as the original essence of life on earth in an “RNA world” but also makes it a suitable candidate for carrying out versatile biological functions (Cech, 2012; Gilbert, 1986).

1.1.1 Classes of eukaryotic RNA

As mentioned before, RNA molecules can provide instructions for protein synthesis. There are, however, other classes of RNA that do not get translated into proteins. These non-coding RNAs (ncRNAs) are mostly classified based on their function and play essential roles in many cellular processes such as transcription regulation and protein synthesis (Mattick and Makunin, 2006; Eddy, 2001; Kapranov et al., 2007). These are the prevalent RNA categories that are relevant for this work:

- **Messenger RNAs (mRNAs)** provide the instructions for ribosomes in the process of protein synthesis (Jackson et al., 2010). Eukaryotic mRNAs may contain untranslated regions (introns) that are spliced out in their maturation process. Splicing is discussed in more detail in section 1.1.2 (Green, 1986).
- **Ribosomal RNAs (rRNAs)** form the most abundant class of RNA molecules, comprising 80% of cellular RNA mass. In eukaryotes four rRNAs – transcribed from two rRNA genes and subsequently processed to form four mature rRNA fragments – bind 79 proteins to form the two ribosomal subunits. The peptidyl-transferase reaction of the ribosome is catalyzed by one of its rRNA molecules, highlighting the role of rRNAs both as enzymes and as structural components of the ribosomes (Henras et al., 2015; Moss et al., 2007). Ribosome production consumes the majority cellular energy and takes up vast nuclear space (Warner, 1999; Pederson, 2011).
- **Transfer RNAs (tRNAs)** connect the mRNA template to the newly synthesizing polypeptide chain by mapping each nucleotide triplet (codon) to its respective amino acid. tRNA-specific aminoacyl-tRNA-synthases "load" tRNAs with their corresponding amino acids, preparing them to enter the translation machinery. 20 ancient well-conserved tRNA aminoacyl-tRNA-synthases exist for each amino acid in the genetic code (Sprinzl et al., 1998; Cusack, 1997; Lodish et al., 2000). The availability of tRNAs influences the speed of protein synthesis. Since the concentration of tRNAs varies in the cell, codon frequencies largely influence translation elongation speed and consequently the amount of cellular proteins (Hanson and Coller, 2018; Bazzini et al., 2016).
- **Small nuclear RNAs (snRNAs)** are a class of short RNA molecules (around 150 nucleotides) primarily involved in mRNA preprocessing (Matera et al., 2007). snRNAs act as scaffolds to attract a specific set of RBPs and form larger complexes called small nuclear ribonucleoproteins (snRNP). snRNP complexes primarily act in various stages of RNA splicing. Therefore, snRNAs encompass both an enzymatic and a structural role, similar to rRNAs discussed before (Kiss, 2004; Will and Lührmann, 2011; Madhani, 2013).
- **Small nucleolar RNA (snoRNAs)** are small RNA molecules that bind RNA modification enzymes and facilitate identification of target RNAs (primarily rRNAs, tRNAs and snRNAs) (Bachellerie et al., 2002; Matera et al., 2007). Their association with their protein partners is specific and the resulting RNA-protein complexes are called small nucle-

olar ribonucleoprotein particles (snoRNPs). snoRNPs identify their target RNA molecules based on their sequence complementarity with the snoRNA in the complex (Kiss-László et al., 1998; Decatur and Fournier, 2003).

- **Cryptic unstable transcripts (CUTs)** were identified by studying newly synthesized transcripts that resulted in the observation that many accessible intra- and intergenic regions get transcribed to produce relatively short RNA molecules (200 to 800 nucleotides), which are quickly removed from the cell (Wyers et al., 2005; Neil et al., 2009; Arigo et al., 2006). These CUTs are often produced by RNA polymerase II (Pol II) that binds at the promoter and transcribes in the opposite direction of the coding transcription unit (Neil et al., 2009). While many CUTs are thought to be by-products of transcription, some have been shown to play a role in gene regulation pathways (Berretta et al., 2008; Martens et al., 2004; Uhler et al., 2007).
- **Stable unannotated transcripts (SUTs)** share many similarities with CUTs: they originate from accessible intra- and intergenic regions and often emerge from protein-coding genomic segments (Marquardt et al., 2011; Xu et al., 2009). However, they have a higher half-life by escaping the immediate targeting by nuclear RNA degradation machinery (Xu et al., 2009).
- **Nrd1-unterminated transcripts (NUTs)** are ncRNAs that describe pervasive transcripts enriched upon depletion of RNA degradation factor, Nrd1. These transcripts have significant overlaps with CUTs and SUTs as Nrd1 can be involved in their nuclear degradation pathway (Schulz et al., 2013; Fox et al., 2015).

1.1.2 RNAs are rarely naked: dynamic RNA-protein interactions regulate the fate of mRNAs

Since mRNAs transfer the genetic information from DNA to proteins, their location and abundance in the cell is tightly controlled to adjust local protein concentrations that in turn determine the cellular phenotype. Taking a closer look at various classes of ncRNAs in the previous section, it becomes evident that most either contribute (directly or indirectly) to mRNA maturation, or play a role in transcription regulation by controlling the speed of mRNAs production. In the following and for the majority of this work, I will focus on the mRNA processing and RNA-protein interactions that involve mRNAs. The major steps of RNA metabolism are described below (Figure 1.2).

RNA transcription

Different cell types are formed in multi-cellular organisms by switching on and off certain genes at developmental checkpoints (Cramer, 2019). Even unicellular eukaryotes such as *Saccharomyces cerevisiae* (budding yeast) require intricate transcription regulation to respond to various environmental stimuli as well as for their growth and development (Lackner et al., 2012; Broach,

2012). mRNA transcription is carried out by Pol II and is largely controlled by selective recruitment of the polymerase to a control region at the beginning of the gene (promoter element) (He et al., 2013). This recruitment can be facilitated by transcription factors (TFs) that bind enhancer elements (termed upstream activation sequences or UAS in yeast) in a sequence-specific manner (Lambert et al., 2018). TFs can boost transcription by recruiting the transcription machinery through cooperative low affinity interactions in their disordered regions (Hahn, 2018; Boija et al., 2018; Ptashne and Gann, 1997). The way this transcription activation is encoded in the disordered regions of TFs is the subject of a collaborative study that I will introduce in chapter 4.3 (Erijman et al., 2020).

Upon assembly of the transcription initiation machinery at the promoter region, the double-stranded DNA becomes unwinded and serves as a template to create the complementary RNA molecule as Pol II marches forward (Cramer, 2019). The growing nascent mRNA chain is co-transcriptionally modified by capping, splicing, cleavage, and polyadenylation complexes (Figure 1.2)(Bentley, 2014).

RNA capping

Capping is the first step in RNA maturation in which the capping enzyme adds a methylated guanosine to the nascent RNA with an unprocessed 5′-triphosphate end (Ramanathan et al., 2016). Capping occurs shortly after the start of transcription and as early as upon the synthesis of the first 20 nucleotides (Martinez-Rucobo et al., 2015). 5′ capping ensures that the nascent transcribing mRNA is protected from the degradation machinery (Jiao et al., 2010). Once capped, the 5′ RNA end is bound by the cap-binding complex (Gonatopoulos-Pournatzis and Cowling, 2014). This complex plays a crucial role in recruiting the necessary factors to the precursor mRNA (pre-mRNA) for spliceosome assembly (Görnemann et al., 2005; Flaherty et al., 1997), polyadenylation (Flaherty et al., 1997), and finally nuclear transport (Cheng et al., 2006; Izaurralde et al., 1995). In the cytoplasm, the mRNA cap recruits initiation factors needed for protein synthesis and helps form the 5′ to 3′ RNA loop during translation which facilitates efficient reinitiaton to enable multiple translation rounds (Fortes et al., 2000; Choe et al., 2012; Vicens et al., 2018).

RNA splicing

Most eukaryotic pre-mRNAs include both non-coding sequences (introns) and protein coding fragments (exons). During splicing introns are removed and the exons are ligated together by the spliceosome complex. The number of introns vastly varies in the eukaryotic kingdom with just few hundred introns in the yeast genome to an average of eight introns per gene in human (Neuvéglise et al., 2011; Sakharkar et al., 2005). The spliceosome is a ribonucleoprotein complex that binds exon boundaries and brings them in close spatial proximity to perform the excision reaction in a step-wise manner (Matera and Wang, 2014). The intron structure is evolutionary conserved and consists of GU and AG dinucleotides that mark the 5′ and 3′ intron boundaries

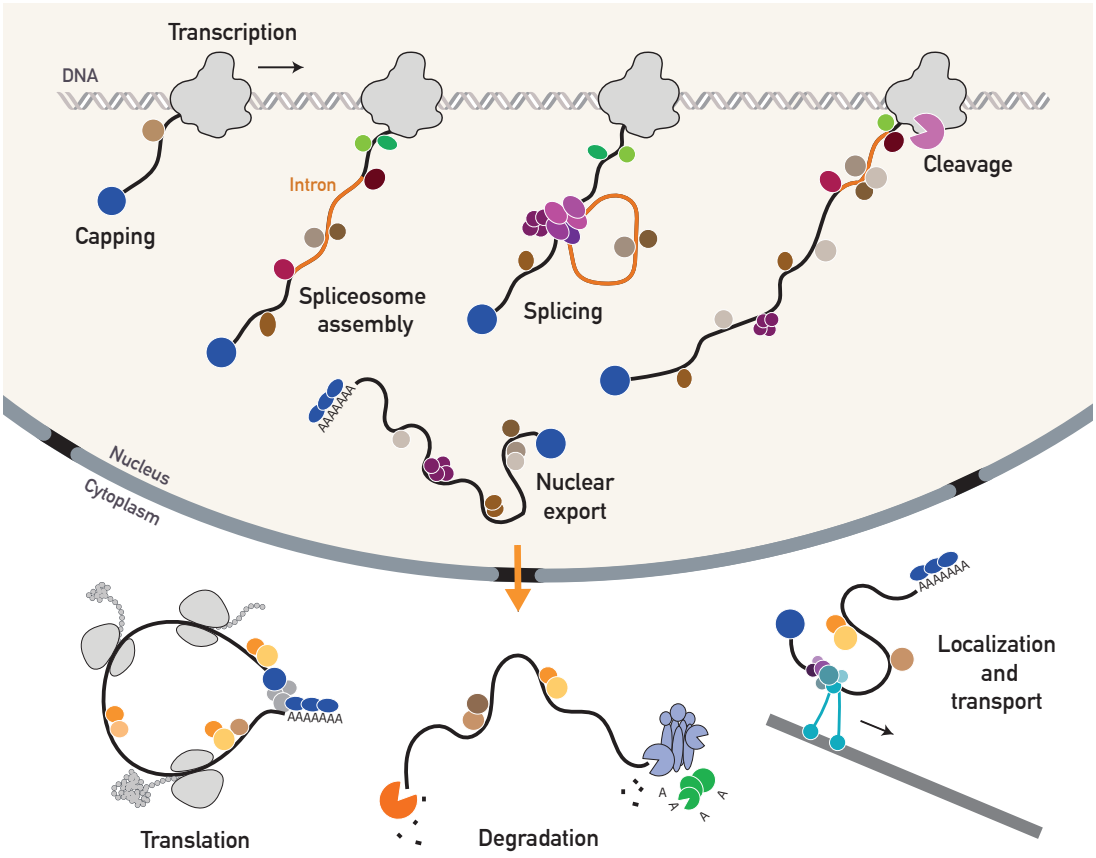


Figure 1.2: **RBPs dynamically interact with mRNAs to guide various stages of RNA processing.** The life of a eukaryotic mRNA starts with transcription in the nucleus by RNA polymerase II. A methylated guanoside is added to the RNA molecule co-transcriptionally by the capping complex. Furthermore, introns are spliced out by the spliceosome complex that, guided by sequence elements in introns, performs the cleavage and ligation reactions. As a last step of pre-mRNA maturation, a poly(A) tail is appended to the RNA molecule. Mature mRNAs are transported to the cytoplasm through nuclear pores by nuclear transport machinery. To control the rate and location of protein production, cytosolic mRNAs dynamically interact with RBPs that transfer them to specific cellular locations, recruit degradation enzymes, or the translation machinery. Adapted from an illustration by Julian König (Buchmann Institute's web page).

respectively, as well as the branch point sequence 18-40 nucleotide sequence upstream of the 3′ splice site. In higher eukaryotes a polypyrimidine tract between the branch point and 3′ intron boundary helps recruit the spliceosome to the 3′ splice site (Herzel et al., 2017; Will and Lührmann, 2011). Various sequence elements in the RNA can act to recruit splicing factors that activate or suppress different steps of the splicing reaction. These splicing enhancer and silencing elements control the fate of introns and regulate alternative splicing, a process that allows a single gene to generate multiple mRNAs by joining together different exon combinations (Matlin et al., 2005; Matera and Wang, 2014).

3′ end cleavage and polyadenylation

The 3′ end of the mRNA is defined by endonucleolytic cleavage. 3′ end cleavage is performed by the termination complex, and is followed by the addition of a long stretch of untemplated adenosines, termed the poly(A) tail, by poly(A) polymerase (Elkon et al., 2013). The recruitment of the termination machinery is controlled by specific motifs (particularly a conserved AAUAAA sequence called polyadenylation signal or PAS) that reside in the 3′ untranslated region of the nascent RNA (Porrua and Libri, 2015; Proudfoot, 2011). Genes in higher eukaryotes often have multiple PASs that can be recognized by the termination machinery, resulting in great variation in the lengths of mRNA molecules produced from a single gene (Elkon et al., 2013; Gruber and Zavolan, 2019). The resulting mRNAs can vary in coding sequences, as well as in their 3′ untranslated region (3′ UTR). Since the 3′ UTRs serve as docking points for many RBPs that regulate RNA function, the variation in 3′ UTR length serves as a regulatory step to control the function of the mRNA as well as its stability, cellular localization, and translation efficiency (Hoque et al., 2013; Lianoglou et al., 2013; Gruber and Zavolan, 2019).

RNA modification

Recent studies have elucidated that mRNAs undergo sequence-specific chemical modifications that can create a binding surface for RBPs or change the RNA structure and flexibility. The bound RBPs can in turn regulate a variety of molecular processes, such as transcription, pre-mRNA splicing, RNA export, mRNA translation, and RNA degradation (Boo and Kim, 2020; Shi et al., 2019). RNA modifications can be dynamic and occur in various stages of the RNA metabolism, both in the cytoplasm and the nucleus (Gilbert et al., 2016).

RNA export

In eukaryotes, RNA transcription and preprocessing takes place in the nucleus, while mRNA translation happens in the cytoplasm. To reach the translation machinery, eukaryotic mRNAs therefore have to pass through the nuclear pore complex (NPC), which tightly regulates the flow of material between the two cellular compartments (Pemberton and Paschal, 2005). NPCs achieve this selectivity by forming a hydrophobic liquid-like mesh, made of phenylalanine-glycine repeats in their disordered regions. They are therefore permeable to particles coated in amino acids that can dissolve in the pore, providing one of the first discovered instances of condensation in cellular biology (Frey et al., 2006; Schmidt and Görlich, 2015). Consequently, nuclear mRNA export is based on the formation of a messenger ribonucleoprotein (mRNP) export complex in the nucleus that is able to diffuse back and forth through the nuclear pore (Stewart, 2010). Transport directionality is imposed by an active process that remodels the mRNP in the cytoplasm and therefore removes key NPC-soluble transport proteins, preventing mRNA return to the nucleus. The RNA export complex is assembled in a step-by-step process that ensures only mature mRNAs (that have undergone capping, splicing and polyadenylation) can exit the nucleus and

reach the translation machinery (Stewart, 2019).

RNA localization and transport

Controlling the location of mRNAs in the cell is an effective way to dictate protein localization and in turn to control cellular function and morphology (Eliscovich and Singer, 2017). A well studied example of RNA localization is the neural mRNA transport from the cell body (where they are transcribed) and across axons to synapses that are sometimes meters away. This synaptic RNA localization allows for rapid changes in protein concentration through on-demand translation of proteins, and is more energy-efficient than transporting many translated proteins from the same mRNA molecule across the axons (Van Driesche and Martin, 2018). The active and directional transport of mRNA in the cytoplasm is facilitated by RBPs that target RNAs in a sequence and structure dependent manner. These RBPs directly or indirectly interact with motor proteins (i.e. kinesins, dyneins and myosins) which transport the mRNP across the cytoskeleton (Gagnon and Mowry, 2011).

Liquid-liquid phase separation (LLPS or condensation) is a newly appreciated concept in biology that explains how some mRNA molecules can get localized without the involvement of motor proteins (Langdon and Gladfelter, 2018). LLPS describes the process in which upon reaching a certain polymer concentration (DNA, RNA, or proteins), the homogeneous solution demixes into a condensed phase (high polymer concentration) and a dilute phase (low polymer concentration). Through demixing the overall number of favorable interactions in a mixture solution increases (Brangwynne et al., 2015; Boeynaems et al., 2018; Banani et al., 2017; Flory, 1942). A well studied example of LLPS-mediated RNA localization is the local assembly of germline RNA granules, termed P granules, during the polarization of *C. elegans*. This process ensures that germplasm components stay exclusively in the posterior side of the embryo and can develop the germline upon cell division (Smith et al., 2016; Brangwynne et al., 2009). P granule localization is thought to be established by the concentration gradients of MEX-5 and -6 that compete with P granule proteins for binding RNA molecules and hence inhibit their RNA-dependent phase separation in the anterior (Seydoux, 2018).

With our growing knowledge of biological pathways that are influenced by LLPS, more examples of localized condensation have been discovered, such as the assembly of microtubules in centrosomes (Conduit et al., 2014), the formation of signalling clusters at the membrane (Case et al., 2019; Banjade and Rosen, 2014), cluster formation at presynaptic active zones (Milovanovic et al., 2018; Zeng et al., 2018), and transciption condensates (Cho et al., 2018; Sabari et al., 2018; Boehning et al., 2018). I will briefly discuss size and localization control of biological condensates in chapter 4.2.

Translation

Translation takes place in the cytoplasm and consists of three major steps: initiation, elongation, and termination. It is a cyclic process, meaning terminated ribosomes are recycled to start a new round of protein synthesis. To enhance the efficacy of this recycling, translating mRNAs often form loops with the help of protein complexes that connect their 3′ and 5′ ends (Vicens et al., 2018; Wells et al., 1998). Regulation of protein synthesis rates is mainly controlled through the first translation step: initiation. The binding of RBPs to the 5′ UTR of the RNA molecule can for example inhibit translation initiation by forming an RNA loop with the cap and blocking the loading of the ribosome. Similarly, other RBPs or microRNAs (miRNAs) that target certain sequences in the transcript can facilitate or hinder translation initiation (Babitzke et al., 2009; Jackson et al., 2010; Muckenthaler et al., 1998).

RNA quality control and degradation

Throughout the previous steps, defective RNA molecules such as those that lack a cap or poly(A) tail, have splicing defects, or carry transcriptional errors resulting in nonsense codons must get recognized and removed from the cell. Degradation of these mRNAs relies on their identification by surveillance RBPs and the subsequent recruitment of the RNA degradation machinery (Doma and Parker, 2007). In addition to removing defective RNAs, functional RNA molecules also undergo regulated degradation to control their concentration and to ensure their removal when they are no longer needed by the cell (Ross, 1996; Miller et al., 2011). Deciphering the RNA-protein interaction landscape in the context of RNA degradation is the topic of the first part of this work. I will therefore introduce this pathway in more depth below.

1.1.3 RNA degradation pathway: An example of harmonious RNA-protein interactions

RNA turnover controls the fate of all eukaryotic RNAs. In yeast, the RNA degradation pathway has four main functions. **(1)** The maturation of many ncRNAs such as snRNAs, snoRNAs, and 5.8S rRNA involves the degradation machinery (Allmang et al., 1999a; Lardelli and Lykke-Andersen, 2020). **(2)** Many quality control pathways are in place to quickly degrade erroneous tRNAs, rRNAs, or mRNAs that would otherwise produce non-functional proteins (He and Jacobson, 2015; Houseley et al., 2006). **(3)** Non-productive Pol II transcripts such as CUTs and NUTs, as well as by-products of RNA preprocessing such as intron fragments are degraded (Wyers et al., 2005; Doma and Parker, 2007). **(4)** The rate at which RNAs get degraded controls RNA abundance in the cell and thereby regulates their function (Ross, 1996; Miller et al., 2011)(Figure 1.3).

In the following I will review the main steps of RNA degradation in yeast, focusing on the players that are studied in the first part of this work.

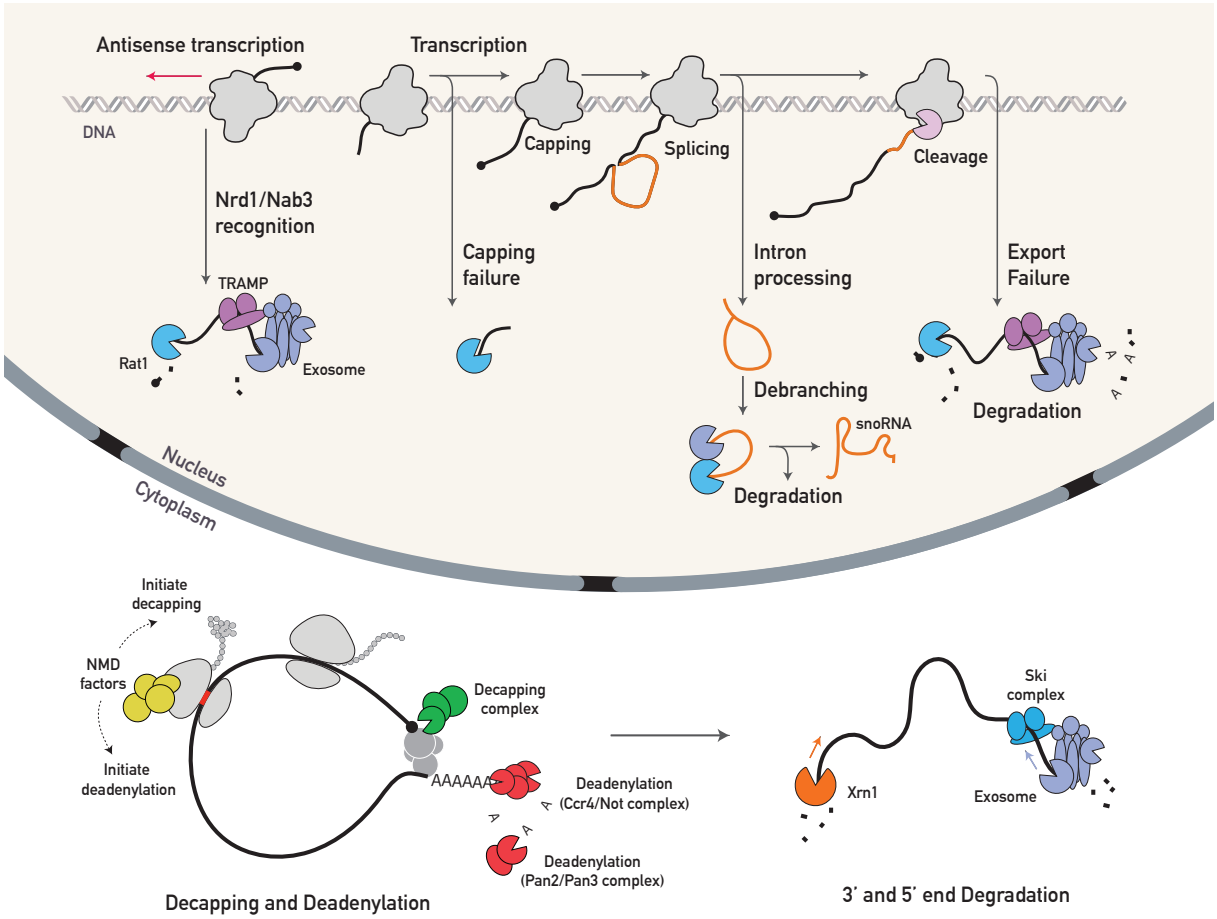


Figure 1.3: **The many pathways of RNA quality control and degradation in yeast**. RNA metabolism is tightly controlled to identify and remove erroneous transcripts quickly. Failure in capping, splicing and polyadenylation can lead to nuclear RNA degradation with the help of 5′ and 3′ exonucleases Rat1 and the exosome complex (in blue). When needed to expose the RNA to exonucleases the cap and poly(A) tail are removed by decapping (in green) and deadenylation complexes (in red) respectively. Transcripts resulting from antisense transcription (a by-product of normal transcription) are targeted by Nrd1/Nab3 and delivered to the degradation machinery. The RNA degradation machinery is also involved in the degradation of spliced introns and preprocessing of many non-coding RNAs such as sn- and sno-RNAs, some of which reside in intronic regions. mRNA degradation also occurs in the cytoplasm in response to translation difficulties (often identified by the nonsense mediated decay machinery, in yellow) or as a means to regulate RNA half-life. Degradation can be triggered by removing the cap or poly(A) tail, causing the opening of the translation loop, and triggering 5′ degradation by the cytosolic exonuclease Xrn1 and/or exosome degradation from the 3′ end.

Degradation initiation by deadenylation and decapping

As mentioned in section 1.1.2, mRNA capping and polyadenylation is a crucial step in RNA maturation. These two end modifications mark the RNA as mature and recruit the cap binding complex (CBC) and the poly(A) binding proteins (PABs) to the two transcript ends to shield it against degradation enzymes. This provides the cell with many ways to control the fate of mRNA through RBPs that stabilize or remove the CBC and PABs. Removing these protective protein complexes would expose the naked ends of the RNA molecule to decapping (removal of

the 5′ cap) and deadenylation (shortening of the 3′ poly-adenylated tail) enzymes, which leads to the subsequent degradation of the mRNA molecule (Parker, 2012).

RNA deadenylation is mainly performed by two complexes: the Ccr4/Not complex consisting of Ccr4, Pop1, Not1-5, and Caf40, and the Pan2/Pan3 complex. Ccr4, Pop1, and Pan3 are the active exonucleases in these complexes and their action is regulated by other associated proteins. The Pan2/Pan3 complex is recruited by Pab1 (a PAB) during mRNA maturation to trim down the size of the poly(A) tail to 70–90 nucleotides (Dunn et al., 2005; Brown and Sachs, 1998). Pab1's presence on the poly(A) tail promotes its trimming by Pan2 while it inhibits the action of Ccr4. This is consistent with a two step model for RNA deadenylation in which Pan2 initiates the trimming to ∼65 residues and then Ccr4 further shortens the poly(A) tail (Parker, 2012; Brown and Sachs, 1998; Tucker et al., 2002).

Decapping in yeast is carried out by the Dcp1/Dcp2/Dcs1 complex with Dcp2 as the catalytically active subunit. (van Dijk et al., 2002; Steiger et al., 2003). Decapping is further regulated by a number of decapping enhancers such as Edc2, Edc3, and Dhh1 that can recruit the decapping complex to initiate the 5′ degradation of mRNA, upon various cellular triggers such as ribosome stalling (Coller and Parker, 2005; Carroll et al., 2011; He et al., 2018).

5′ to 3′ mRNA degradation

Once the cap structure is removed, the mRNA's 5′ monophosphate is prone to 5′ → 3′ degradation by the exonuclease Xrn1 (Jinek et al., 2011; Stevens, 2001). Xrn1 couples its processing with unwinding of local RNA structures, making it independent of helicases (Jinek et al., 2011; Parker, 2012). Xrn1 has a paralog, Rat1, which is localized in the nucleus and is involved in nuclear 5′ → 3′ RNA degradation and preprocessing (Park et al., 2015; Schmid and Jensen, 2018; Baejen et al., 2017).

3′ to 5′ mRNA degradation

Upon sufficient shortening of the Poly(A) tail, further 3′ → 5′ degradation of the RNA is carried out by the exosome and its associated factors. The core exosome consists of 10 subunits: the catalytically active exonuclease Rrp44 (can also perform endonucleation), together with three small RBPs (Rrp4, Rrp40, and Csl4), as well as six members of the RNase PH protein family (Rrp41, Rrp42, Rrp43, Rrp45, Rrp46, and Mtr3) (Allmang et al., 1999b; Park et al., 2015; Liu et al., 2006). The first step of RNA degradation by the exosome is the passage (and identification) of RNA through the TRAMP complex (in nucleus) or Ski complex (in cytoplasm) (Houseley and Tollervey, 2009; Park et al., 2015). The TRAMP complex is involved in many nuclear preprocessing and quality control mechanisms and exists in two isoforms: TRAMP4 (Trf4, Air2 and Mtr4) and TRAMP5 (Trf5, Air1 and Mtr4) (Anderson and Wang, 2009; Houseley and Tollervey, 2008). It harbors a poly-(A) polymerase (Trf4 or Trf5) thought to make the RNA substrate more attractive for exonucleation (Jia et al., 2011; Vaňáčová et al., 2005; LaCava

et al., 2005), a zinc-knuckle putative RBP responsible for RNA recognition (Air1 or Air2), and an RNA helicase (Mtr4) (Hamill et al., 2010; Falk et al., 2014). A more recent study suggests a third isoform consisting of Trf4, Air1 and Mtr4 (Delan-Forino et al., 2020). Nuclear exosome additionally associates with the 3′ exonuclease Rrp6 that takes part in antisense RNA decay and aberrant mRNA degradation (Callahan and Butler, 2010; Davis and Ares, 2006; Danin-Kreiselman et al., 2003). The Ski complex accompanies the exosome for cytosolic RNA degradation. It consists of Ski2, an RNA helicase, as well as Ski3, and Ski8 (Brown et al., 2000; Wang et al., 2005). Ski7 and Ski4 have been reported to bind cytosolic exosome directly (van Hoof et al., 2002). Both TRAMP and Ski complexes contribute to substrate specificity by reading out various degradation signals, such as Nrd1/Nab3 mediated recognition of aberrant transcripts in the nucleus or ribosome mediated translation difficulties in the cytosol (Schmidt and Butler, 2013a; Delan-Forino et al., 2020; Schmidt and Butler, 2013b).

Nuclear surveillance and preprocessing of ncRNAs

In addition to regulating the quality and stability of mRNAs, the nuclear degradation machinery is involved in the maturation of pre-snRNAs, pre-snoRNAs, pre-tRNAs, and pre-rRNAs through trimming and cleavage. Moreover, the spacer fragments produced during rRNA biogenesis as well as non-functional introns are removed (Allmang et al., 1999a). Furthermore, the degradation machinery helps remove CUTs, NUTs, SUTs, and aberrant ncRNAs and mRNAs through communication with the surveillance pathway (Sloan et al., 2012; Thiebaut et al., 2006).

1.2 How do proteins target specific RNA molecules?

A common thread between all processes described in the last sections is the dynamic involvement of RBPs in each step of RNA biochemistry. These RNA-protein interactions control the fate of mRNA molecules by regulating their transcription, stability, cellular location, and translation rates (Singh et al., 2015; Dreyfuss et al., 2002). RNA molecules can also regulate RBP function by altering their stability, interaction partners, and localization (Hentze et al., 2018). Recent estimates suggest that the human genome may encode for more than 1500 RBPs (encompassing 7.5% of all protein-coding genes), highlighting the importance of RBPs (Gerstberger et al., 2014). To ensure that the correct RNA molecules are targeted, RBPs must bind with high specificity. I will briefly describe four major aspects of obtaining RBP specificity (Figure 1.4): **(1)** selecting specific RNA sequences and structures, **(2)** cooperative multi-domain binding, **(3)** cooperativity among various RBPs, and **(4)** co-localization through condensate formation.

1.2.1 Selecting specific RNA sequences and structures

RBPs often bind RNA using various structured RNA-binding domains (RBDs) (Castello et al., 2016; Lunde et al., 2007) or sometimes also with disordered regions such as RGG/RG and RS

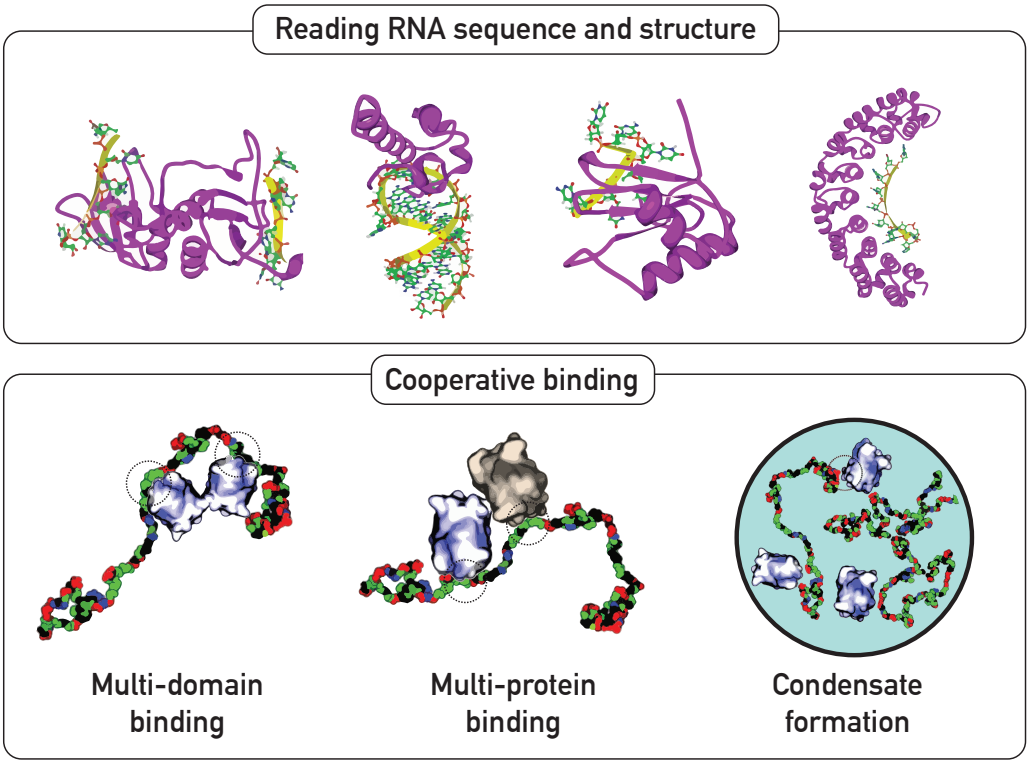


Figure 1.4: **RBPs find their target through a mixture of sequence and structural specificity and cooperative binding. (Top)** Many RBPs have RBDs or unstructured sequence elements with an affinity towards specific RNA sequences and/or structures. Examples are from left to right: PTB, binding domains 3 and 4, in complex with CUCUCU RNA [Protein Data Bank (PDB): 2ADC], Vts1p sterile-α motif domain in complex with a 5′-CUGGC-3′ pentaloop embedded in a 19nt hairpin [PDB: 2ESE], RBD1 of PTB in complex with CUCUCU RNA [PDB: 2AD9], and structure of the Pum1 PUM-homology domain in complex with the single-stranded RNA 5′-AUUGUACAUA-3′. This structure demonstrates an extreme case with high sequence specificity as the last 8 nucleotides are individually recognized by 8 Puf repeats in the PUM domain [PDB: 1M8Y]. Visualizations of RNA-protein structures are taken from Li et al. **(Bottom)** Higher levels of specificity can be achieved by stacking several RBDs, or by favoring interactions in the presence of multiple RBPs. This can be either due to protein complex formation or a result of transient interactions between disordered regions of these proteins. Higher concentrations of RNA and proteins resulting from condensate formation can further boost affinity and specificity of RBP-RNA interactions. Illustrations are adapted from Pak et al.

motifs which are known to modulate RNA-binding activity (Ozdilek et al., 2017; Calabretta and Richard, 2015). These RBDs can engage with specific RNA sequences and structures. For instance many RNA-recognition motifs (RRMs) recognize single-stranded bases specifically through the protein β-sheet and two loops that connect the secondary structure elements (Figure 1.4, top)(Oberstrass et al., 2005; Lunde et al., 2007). While RRMs in different proteins fold into a similar structure, small variations in the amino acid residues in critical positions can give rise to RBPs that recognize different RNA sequences.

Unlike TFs that target genomic sequences 6-12 nucleotides in length (Lambert et al., 2018), RBDs often recognize very short sequences ($\sim$3 nucleotides and rarely above 5)(Ray et al., 2013; Dominguez et al., 2018). RBPs can partially compensate for this by adopting cooperative

binding (described below) as well as spatial and temporal control of RNA and protein abundance.

1.2.2 Multi-domain binding

RBPs are often modular, consisting of multiple RBDs. A closer look at putative human RBPs shows that more than half contain multiple RBDs of distinct types (Figure 1.5). Multiple domains allow the protein to recognize longer stretches of RNAs or sequences that are separated from each other on the RNA (Lunde et al., 2007). Higher affinities can be achieved by cooperative binding due to an increased local concentration of the RNA molecule at an unbound domain when another domain is already bound. We have shown that this effect results in dissociation constants (K_d) for the multi-domain RBP that can be several orders of magnitude smaller than that of each domain in the protein (Stitzinger et al., 2021). I will introduce this study in section 4.1.

A well-studied example of multi-domain binding is the mRNA-binding protein IMP3, which contains six RBDs: four K-homology (KH) and two RNA-recognition motif (RRM) domains. Studying RNA fragments that are bound by IMP3 has shown that all domains contribute to the overall specificity. Consequently, IMP3 identifies appropriately spaced CA-rich and GGC-core RNA elements, that can span over a hundred nucleotides (Schneider et al., 2019). Others have also reported evidence for spaced sequence preferences in about one third of the studied RBPs, highlighting the importance of multi-domain binding in modulating RBP specificity (Dominguez et al., 2018; Jolma et al., 2020).

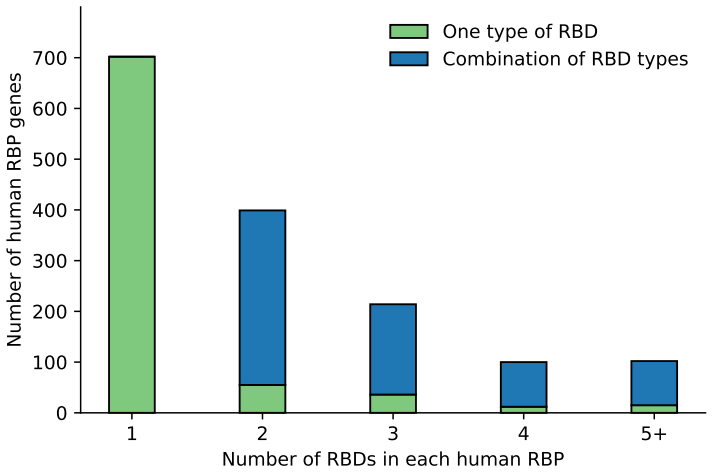


Figure 1.5: **Many RBPs have multiple RBDs of various types.** This graph shows the number of RBDs annotated in each RBP gene in human. Proteins that harbor various types of RBDs (such as RRMs, KH-domains, and ZFs) are marked in blue. RBPs with only one domain or repeats of the same domain type are marked in green. The data used to generate this plot is taken from Gerstberger et al..

1.2.3 Cooperative binding among multiple RBPs

Similarly to increasing affinity and specificity by stacking RBDs, proteins can bind target RNA substrates cooperatively through protein-protein interactions either in interaction domains or their disordered regions. As seen in previous chapters (1.1.2 and 1.1.3) many proteins interact to form stable complexes such as those involved in splicing, decapping, deadenylation, and exonucleation. A well-studied example is the Nrd1/Nab3 complex in yeast. Both Nrd1 and Nab3 bind RNA molecules that target GUAG and CUUG RNA sequences respectively (Sohrabi-Jahromi et al., 2019; Schulz et al., 2013). A small difference in density of these two motifs between the sense (gene-coding) and antisense (opposite) strand are sufficient for Nrd1/Nab3 targeting of aberrant transcripts and their subsequent RNA degradation (Schulz et al., 2013).

1.2.4 Co-localization in biological condensates

As explained in section 1.1.2, condensation can lead to higher local concentrations of RNA and protein molecules, resulting in an increase in their interaction probabilities. Examples of well characterized ribonucleoprotein granules are: nucleoli (Brangwynne et al., 2011), transcriptional condensates (Cho et al., 2018), nuclear speckles (Galganski et al., 2017), Cajal bodies (Sawyer et al., 2017), processing bodies (P-bodies) (Teixeira and Parker, 2007), stress granules (SGs) (Molliex et al., 2015), and germ granules (Smith et al., 2016). For example highly cooperative interactions among the C-terminal domain (CTD) of the transcribing Pol II, its nascent RNA product and several other nuclear proteins can lead to condensate formation (Boehning et al., 2018; Sabari et al., 2018; Cho et al., 2018). These transcription condensates can efficiently recruit the RNA preprocessing machinery to facilitate pre-mRNA maturation (Guo et al., 2019; Cramer, 2019). Similarly rRNA transcription stabilizes nucleoli, a sub-nuclear compartment specialized for ribosome biogenesis (Berry et al., 2015; Feric et al., 2016). Interestingly, rRNA transcription, processing, and assembly into pre-ribosomes all occur within three distinct membraneless compartments within the nucleolus. This intricate organization ensures that the process is efficient and the steps are followed in the desired order (Sabari et al., 2020; Pederson, 2011; Strom and Brangwynne, 2019).

1.3 Experimental and computational approaches to uncovering RBP specificity

While structural determination of RNA-protein complexes and biochemical assays for studying the dynamics of these interactions have been instrumental for understanding the chemistry of protein-RNA interactions, advances in high-throughput sequencing technologies set a milestone by enabling the identification of global RBP binding sites inside living cells or in test tubes. The availability and affordability of high-throughput sequencing has resulted in the development of dozens of experimental protocols for studying RNA-protein interactions and petabytes of

sequencing data to explore with computational methods. In the following sections, I will first introduce commonly used high-throughput sequencing technologies, and then summarize current computational approaches for modeling RNA-protein interactions.

1.3.1 Uncovering protein binding sites with high-throughput sequencing technologies

Several experimental techniques have emerged to obtain systematic maps of RBP binding sites *in vivo* (Hentze et al., 2018). These approaches are often based on RNA immunoprecipitation and subsequent sequencing (RIP-seq)(Gilbert and Svejstrup, 2006). Here, RNA fragments that are bound to an immunoprecipitated protein of interest are purified. The bound RNA fragments are then sequenced and mapped to the genome. Binding regions are then identified based on statistical evaluation of the read profiles (Uhl et al., 2017). A common additional step to this approach is cross-linking the protein to its bound RNA fragment before purification, termed cross-linking immunoprecipitation (CLIP-seq)(Licatalosi et al., 2008). Cross-linking reduces the experimental noise by allowing a more rigorous washing step and grants a higher resolution in identification of the binding sites. Several variations of the CLIP-seq protocol have been developed that can determine the binding footprints with single-nucleotide resolution: photoactivatable-ribonucleoside-enhanced CLIP (PAR-CLIP)(Hafner et al., 2010), individual-nucleotide-resolution CLIP (iCLIP)(König et al., 2010), and enhanced CLIP (eCLIP)(Van Nostrand et al., 2016).

Mapping RBP binding sites *in vivo* is a valuable approach for uncovering the cellular function of the studied protein. However, deriving accurate motif models of RNA-protein interactions from *in vivo* data is challenging due to complications arising from cooperativity and competition with other RBPs (Dominguez et al., 2018), high levels of non-specific background binding (Friedersdorf and Keene, 2014), and the influences of RNA localization, expression, and folding (Änkö and Neugebauer, 2012). Therefore, additional techniques have been developed to study the binding preferences of RBPs *in vitro* and in isolation from other RBPs. These technologies often include the creation of a random pool of RNA sequences, selection of bound fragments by protein immunoprecipitation, and their subsequent identification by sequencing. RNA-compete, the first high-throughput approach, identified bound RNA fragments with microarrays (Ray et al., 2013). However, current approaches rely on high-throughput sequencing. These techniques include RNA-compete-seq (Cook et al., 2017), RNA bind-n-seq (RBNS)(Dominguez et al., 2018), and high-throughput RNA systematic evolution of ligands by exponential enrichment (HTR-SELEX)(Jolma et al., 2020).

The data presented in the first part of this work is generated by PAR-CLIP experiments. The second part of my thesis, primarily focuses on the analysis of HTR-SELEX data. I will therefore introduce these two techniques in more depth here (Figure 1.6).

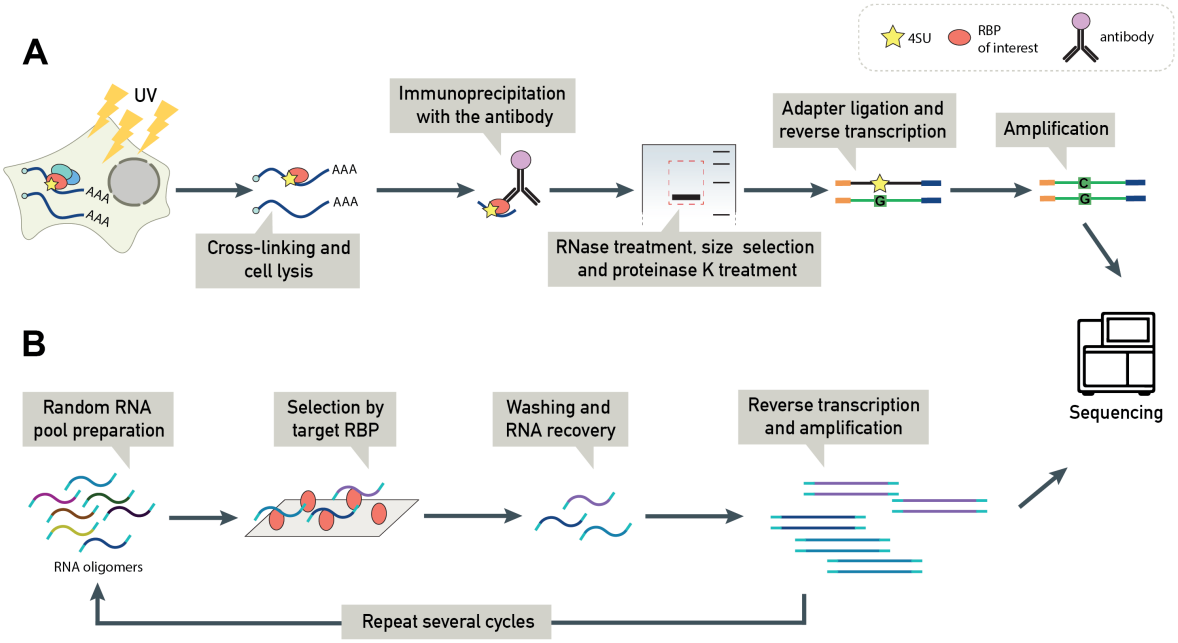


Figure 1.6: **Experimental identification of RBP binding sites (A)** PAR-CLIP protocol. Cells are supplemented with 4SU that incorporates into nascent RNAs as a uridine replacement. They are then exposed to UV radiation which creates 4SU-protein cross-links. After cell lysis, the protein-RNA complexes are purified by immunoprecipitation of the target RBP. After partial digestion and size selection by gel electrophoresis, the final RNA fragments are amplified sequenced, and mapped to the genome resulting in T to C transitions at cross-link positions. Figure is adapted from Hentze et al.. **(B)** HTR-SELEX protocol. A random RNA pool is incubated with the target RBP. Bound RNA fragments are washed from the resin, reverse transcribed and amplified. The RNA fragments are then identified by high-throughput sequencing. This process can be repeated several times to enrich RNA oligomers bound with higher affinity.

PAR-CLIP protocol

PAR-CLIP is the first protocol developed to achieve single nucleotide resolution in determining protein binding sites. Cells are cultured on media that is supplemented with the modified nucleotide 4-thiouridine (4SU). 4SU readily incorporates into nascent RNAs as a uridine replacement. This is followed by UV radiation that leads to the cross-linking of incorporated 4SU nucleotides with interacting RBPs. The cells are then lysed and the RBP of interest is purified through immunoprecipitation with a matching antibody. The RNA molecules are partially digested afterwards with RNase T1 to produce smaller fragments, while bound RNA regions are protected by the cross-linked protein. To ensure that only RNAs bound to the desired protein are sequenced, the cross-linked RNA protein complexes are separated with gel electrophoresis and size selected. The protein is removed after a proteinase K digestion step and the remaining RNA sequences are amplified and then sequenced (Spitzer et al., 2014; Hafner et al., 2010; Garzia et al., 2017).

The cross-linked 4SU will be recognized as cytidine analogs by the reverse transcriptase during complementary DNA (cDNA) library preparation. This results in a thymidine to cytidine

(T→C) transition in PAR-CLIP sequences at the cross-link positions. Downstream processing of the dataset therefore involves the use of statistical methods to identify high-confidence binding sites based on the frequency of T→C transitions (Roth and Torkler, 2019; Corcoran et al., 2011; Comoglio et al., 2015).

HTR-SELEX protocol

To identify RNA sequences that bind a selected protein, HTR-SELEX uses random DNA sequences of a defined length (typically 20 or 40 nucleotides), which contain 5′ and 3′ primer sequences. These sequences are amplified and transcribed into RNA using the viral T7 RNA polymerase. The transcribed random RNA sequences are incubated with the recombinant protein of interest and the protein-bound fragments are purified using chromatographic techniques. After washing, the bound RNA fragments are amplified during the cDNA library preparation and subsequently sequenced. The selection-amplification-sequencing cycle is then repeated to further select for higher affinity binding partners (Jolma et al., 2020; Schneider et al., 2019).

1.3.2 Current approaches to *de novo* RNA motif discovery

Understanding the mRNP code, that is decoding the basis of specificity in cellular RNA-protein interactions, is key to deciphering the RNA regulatory network and to understanding the relationship between the RNA sequence and its function (Gehring et al., 2017; Brannan and Yeo, 2016; Hennig and Sattler, 2015). To reach this goal, a wide range of motif discovery tools have been developed to infer binding models based on the large amount of available *in vivo* and *in vitro* datasets. In the next sections, I will first introduce commonly used models to represent RNA motifs. Next, I will summarize current approaches to learning these motif models. Finally, I will summarize the limitations of existing motif discovery tools.

Motif models

De novo RNA motif discovery entails the search for over-represented patterns in bound RNA sequences that originate from the binding of the target RBP. There are several approaches to modeling RNA motifs (Figure 1.7). The first and simplest approach is to represent the motif with a linear RNA sequence, such as the GUAG motif used to describe the binding of Nrd1 (Schulz et al., 2013; Hashim et al., 2019). The second and most commonly used motif model is the positional weight matrix (PWM). The PWM takes the degeneracy of the sequence model into account by assigning weights for observing each nucleotide at each position. The PWM assumes that the nucleotide probabilities between the positions are independent (Hartmann et al., 2013). Bayesian Markov modeling (BaMM) is a third approach that overcomes this independence assumption by representing the sequence preferences as conditional probabilities, including the dependencies on preceding nucleotides (Siebert and Soeding, 2016; Kiesel et al., 2018). This is an extension of hidden Markov models (HMMs) that only consider the dependencies between

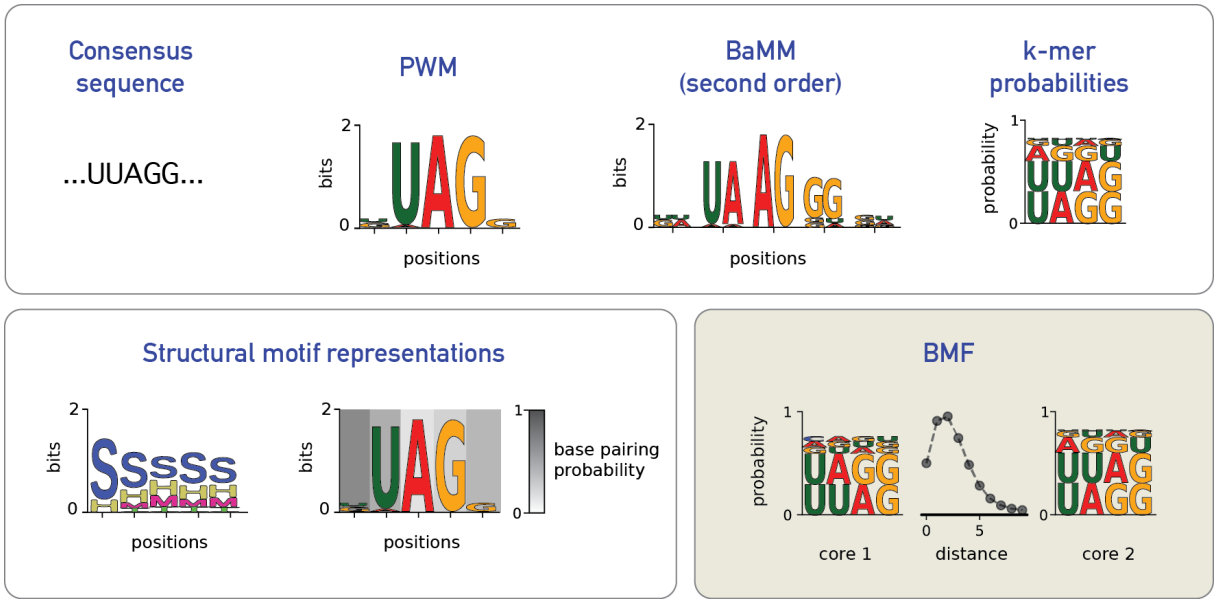


Figure 1.7: **A variety of motif models are used to represent RBP sequence and structural preferences.** Sequence representations include: the linear consensus sequence, PWMs, BaMMs, and k-mer energies (or probabilities). To represent secondary structure an extended alphabet can be used to show preferences for or against common RNA structures, or alternatively base pair probabilities are shown alongside the sequence motif. In the second part of this thesis I will introduce a new RNA motif model learned by Bipartite Motif Finder (BMF) that models bivalent RNA-protein interactions (Sohrabi-Jahromi and Söding, 2021).

adjacent nucleotides (Heller et al., 2017). However, none of these models are able to differentiate between multi-modal (preferences for several distinct sequences), or variably gapped binding specificities (Sasse et al., 2018). A forth type of RNA motifs assigns different binding energies to all RNA oligomers up to a given length k (also known as k-mers). The advantage of this approach is that it allows the learning of not only the inter-positional dependencies but also multi-modal motifs. However, the number of parameters increases exponentially with k (4^k), making such models prone to overfitting when training longer motifs (Schneider et al., 2019; Dominguez et al., 2018; Orenstein et al., 2016; Sasse et al., 2018).

It has been estimated that around a third of RBPs have a preference for binding a certain RNA structure or structural context (Ray et al., 2013; Dominguez et al., 2018; Jolma et al., 2020). The motif models mentioned above can be adapted to include this information. A common approach is the use of an additional alphabet to represent structural preferences in the form of a consensus sequence or PWM. In a simplified version, single-stranded and double-stranded bases are differentiated (e.g. DDDDSSSDDD can represent an RNA stem loop). However, more often this alphabet can be extended to include diverse RNA secondary structures, such as stems (S), multiloops (M), hairpins (H), internal loops (I), bulge (B), and external regions (E) (Pan et al., 2018; Zhang et al., 2016). Another approach for representing structural motifs is to include base pairing probabilities for each nucleotide in the sequence motif model (Munteanu et al., 2018; Jolma et al., 2020).

Motif discovery tools

Many computational methods have been developed to fit the previously mentioned motif models based on sets of motif-enriched (bound to RBP) and background (not bound to RBP) RNA sequences. This is often achieved by fitting models that best distinguish bound and unbound sequences. These approaches can be broadly categorized as explicit and implicit motif finders (Sasse et al., 2018). Explicit motif discovery tools optimize a specific motif model (such as a PWM) that is most enriched in bound sequences (Munteanu et al., 2018; Siebert and Soeding, 2016; Hiller et al., 2006; Bahrami-Samani et al., 2014). Implicit motif detection tools, however, directly learn how to distinguish enriched and background sequences without learning an explicit motif representation. The motifs are then inferred after the binding model is fit and only represent an approximation to the RBP binding preference learned by the model. Examples of this approach are non-linear machine learning methods such as support vector machines (SVMs) (Maticzka et al., 2014) and deep learning approaches such as convolutional neural networks (Alipanahi et al., 2015; Pan et al., 2019; Yan and Zhu, 2020). In fact, many recently developed algorithms use deep neural networks to predict RBP binding sites. These networks often have many parameters that allow them to implicitly learn complex patterns embedded in the dataset, such as RNA structure, multi-modal motifs, co-occurrence of motifs from other RBPs, and experimental biases (Sasse et al., 2018; Ghanbari and Ohler, 2020). Moreover, additional information can be incorporated in the training step to increase the accuracy of model predictions. This information includes: the secondary RNA structure (Pan et al., 2018; Maticzka et al., 2014; Budach and Marsico, 2018; Zhang et al., 2016; Ben-Bassat et al., 2018; Su et al., 2019; Deng et al., 2020), and when using *in vivo* data, the position of binding sites in mRNA (i.e. $3'$ and $5'$ UTRs, coding sequence, or introns), gene annotation, and co-occurrence with other RBPs (Stražar et al., 2016; Pan and Shen, 2017; Yu et al., 2018; Avsec et al., 2017).

Challenges and limitations of current motif discovery approaches

While deep-learning approaches have shown promising accuracy in predicting bound and unbound sequences, the logic behind their classification remains poorly understood (Ghanbari and Ohler, 2020). There have been many efforts to make these "black box" neural networks more interpretable. "Mutation maps" that show the effect of point mutations in the predicted binding score can help highlight sequence regions that are more critical for protein binding (Alipanahi et al., 2015). In the case of convolutional neural networks, filters of the convolutional layers resemble PWMs and can provide an insight on sequence features learned by the network (Ghanbari and Ohler, 2020; Pan and Shen, 2017; Pan et al., 2018; Pan and Shen, 2018). Other approaches rely on tracking the dependencies between input features and neural network predictions through gradient calculations, highlighting features in the sequence that are most important for identifying binding sites (Shrikumar et al., 2016; Sundararajan et al., 2017; Ghanbari and Ohler, 2020; Jha et al., 2020). Further postprocessing steps are needed to convert these feature importance maps to binding motifs. Overall, while the trained network can be used to estimate RBP motifs, the resulting motif model is an estimation and does not reflect the infor-

mation learned by the network one-to-one. Moreover, with an increase in the number of model parameters the risk grows that highly parametric models such as deep neural networks could learn biases in experimental datasets. These experiment-dependent biases can be a result of library preparation, amplification, or can depend on the type and concentration of RNase that is used (Kishore et al., 2011; Orenstein and Shamir, 2014). For instance, the PAR-CLIP protocol detects binding sites based on T to C mutations and upon partial digestion by RNase T1, which cleaves after guanines. This makes it prone to enrich cross-link sites in regions with high thymine and guanine frequencies (Friedersdorf and Keene, 2014; Kishore et al., 2011). Complex models can learn these subtle biases to better distinguish bound from unbound sequences. This overfitting can be particularly problematic for RBPs, since they often bind short and repetitive sequences in low complexity regions of UTRs (Dominguez et al., 2018).

Another challenge for studying specificities of RBPs is that they often bind their substrate highly cooperatively through multi-domain binding or self-association (see Figure 1.5 that shows more than half of RBPs have multiple RBDs)(Lunde et al., 2007; Ray et al., 2013). This results in the enrichment of multiple short sequences in the bound RNA with flexible gap lengths between each pair. The spacing, while partially flexible, is influenced by the structure of the RBP and is specific to each RBP (Schneider et al., 2019; Dominguez et al., 2018; Jolma et al., 2020). Learning the co-occurence of RNA motifs together with their spacing is therefore crucial for understanding RBP targeting and could help explain part of the reported missing specificity for RBPs (Jankowsky and Harris, 2015).

1.4 Motivation and aims of this thesis

RBPs control many aspects of RNA metabolism from synthesis to degradation. Understanding how these proteins work and how they identify their target RNA molecules has been the central focus of my doctoral research. My work revolves around two main topics: **(1)** characterization of RBPs in the context of RNA degradation pathway, and **(2)** developing a thermodynamic model to learn multivalent specificities of RBPs.

1.4.1 Genome-wide characterization of general eukaryotic RNA degradation factors

The last event in the life of RNA molecules is their targeted degradation via the complex RNA degradation machineries. While transcriptome-wide interaction maps had been produced for many RBPs involved in transcription initiation, elongation, termination, surveillance, RNA preprocessing, and nuclear transport in yeast (Schulz et al., 2013; Baejen et al., 2017; Battaglia et al., 2017; Baejen et al., 2014), global maps of many general RNA degradation factors were missing when I started my PhD. In a collaboration with the laboratory of Prof. Patrick Cramer, RBP-RNA interaction maps *in vivo* were generated using the PAR-CLIP protocol. In the first part of this thesis, I report on the analysis of the first RBP interactome dataset of yeast degradation factors to address the following points:

1. What are the key differences between the 5′ and the 3′ degradation machineries? Do they process different substrates?
2. Is there a variation in role and specificity inside each degradation complex?
3. Which degradation complexes are responsible for interacting with transcription and translation surveillance systems?
4. What is the rate limiting step of RNA degradation?

1.4.2 Thermodynamic modeling of multivalent RNA-protein interactions

Many cellular processes rely on the binding of RBPs with high affinity and specificity to a specific subset of RNA molecules in the cell (section 1.1). However, studying RBP specificities has shown that they bind short and degenerate RNA sequences (Dominguez et al., 2018). This so-called "missing specificity" of RBPs is partially attributed to their modular nature, the fact that they contain many RBDs that can target several small RNA fragments simultaneously (Jankowsky and Harris, 2015; Lunde et al., 2007; Schneider et al., 2019; Nicastro et al., 2017). Since the gap between the bound sequences is variable and dependent on the RBP structure, current motif models (summarized in Figure 1.7) are not able to capture and represent multivalent binding. Moreover, to estimate binding affinities to such repetitive and degenerate target sequences, it is important to take the many binding configurations with a similar total binding energy into account. This calls for a thermodynamic approach that can correctly sum up the contributions from all these binding configurations. In the second part of my work, I have developed a computational tool for learning bipartite RNA motifs in RNA-protein interaction datasets, termed bipartite motif finder (BMF). BMF is the first tool that adopts a thermodynamic approach to motif discovery. By developing BMF I aimed at addressing the following questions:

1. Would BMF's bipartite motif models outperform other existing RNA motif discovery tools in predicting new binding sites?
2. How prevalent is bipartite binding among RBPs?
3. How long and complex are sequence preferences of multivalent RNA binders?
4. Can BMF reliably learn the spatial geometry between the protein binding sites?

2 Transcriptome maps of general eukaryotic RNA degradation factors

Publication:

"Transcriptome maps of general eukaryotic RNA degradation factors."

S. Sohrabi-Jahromi*, K.B. Hofmann*, A. Boltendahl, C. Roth, S. Gressel, C. Baejen, J. Söding†, P. Cramer†.

(*) first author, (†) corresponding author

eLife (2019): e47040.

2.1 Author contributions

P. Cramer (PC) and J. Söding (JS) conceptualized the research. **S. Sohrabi-Jahromi (SSJ)** designed and performed all the data analyses except the PAR-CLIP preprocessing pipeline, and created all the visualizations except Figure 1 – figure supplement 2. JS, PC, and K.B. Hofmann (KBH) contributed data analysis ideas. KBH, A. Boltendahl, S. Gressel, and C. Baejen carried out the PAR-CLIP experiments. C. Roth created the preprocessing pipeline and assisted with adapting the pipeline to this PAR-CLIP dataset. The preprocessing pipeline refers to the computational steps that start with the quality control of the raw sequencing data and outcome a list of high-confidence cross-link sites. **SSJ**, KBH, JS, and PC wrote the manuscript with input from all authors.

2.2 Code and data availability

All the analysis scripts developed to produce this work are available at `https://github.com/soedinglab/Degradation_scripts`. Raw and preprocessed sequencing data is deposited in NCBI Gene Expression Omnibus under the accession codes GSE 128312.

eLIFE elifesciences.org

RESEARCH ARTICLE

Transcriptome maps of general eukaryotic RNA degradation factors

Salma Sohrabi-Jahromi[1†], Katharina B Hofmann[2†], Andrea Boltendahl[2], Christian Roth[1], Saskia Gressel[2], Carlo Baejen[2], Johannes Soeding[1*], Patrick Cramer[2*]

[1]Quantitative and Computational Biology, Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany; [2]Department of Molecular Biology, Max-Planck-Institute for Biophysical Chemistry, Göttingen, Germany

Abstract RNA degradation pathways enable RNA processing, the regulation of RNA levels, and the surveillance of aberrant or poorly functional RNAs in cells. Here we provide transcriptome-wide RNA-binding profiles of 30 general RNA degradation factors in the yeast *Saccharomyces cerevisiae*. The profiles reveal the distribution of degradation factors between different RNA classes. They are consistent with the canonical degradation pathway for closed-loop forming mRNAs after deadenylation. Modeling based on mRNA half-lives suggests that most degradation factors bind intact mRNAs, whereas decapping factors are recruited only for mRNA degradation, consistent with decapping being a rate-limiting step. Decapping factors preferentially bind mRNAs with non-optimal codons, consistent with rapid degradation of inefficiently translated mRNAs. Global analysis suggests that the nuclear surveillance machinery, including the complexes Nrd1/Nab3 and TRAMP4, targets aberrant nuclear RNAs and processes snoRNAs.

DOI: https://doi.org/10.7554/eLife.47040.001

***For correspondence:** johannes.soeding@mpibpc.mpg.de (JS); patrick.cramer@mpibpc.mpg.de (PC)

[†]These authors contributed equally to this work

Competing interests: The authors declare that no competing interests exist.

Funding: See page 25

Received: 24 March 2019
Accepted: 27 May 2019
Published: 28 May 2019

Reviewing editor: Torben Heick Jensen, Aarhus University, Denmark

Introduction

The abundance of the different eukaryotic RNA species controls cell type and cell fate, and is determined by the balance between RNA synthesis and RNA degradation. Multiple mechanisms exist for RNA degradation (***Parker, 2012***). RNAs are generally exported to the cytoplasm, where they are degraded with an RNA-specific rate. Such canonical turnover is critical for RNA homeostasis. However, RNAs that are defective with respect to their processing, folding, assembly into RNA-protein particles, or their ability to be translated, are identified and rapidly degraded by surveillance pathways. Surveillance occurs both in the nucleus (***Schmid and Jensen, 2018***) and in the cytoplasm (***Zinder and Lima, 2017***). In the nucleus, aberrant RNAs resulting from upstream antisense transcription are quickly degraded. Moreover, some non-coding RNAs (ncRNAs) require processing by the degradation machinery (***Houseley et al., 2006***). Cytosolic RNAs also vary in their life-time, with mRNAs encoding cell cycle regulators or transcription factors having reported life-times in the range of minutes (***Geisberg et al., 2014***; ***Miller et al., 2011***), whereas ribosomal RNAs live for days (***Turowski and Tollervey, 2015***). Therefore, RNA degradation kinetics need to be actively regulated to ensure the optimal life time for each transcript.

RNA degradation can occur from both ends of a transcript, and these processes are often coupled. During canonical mRNA turnover, degradation is thought to be initiated by shortening of the 3′ poly-adenylated (polyA) tail through two major deadenylation complexes, the Pan2/Pan3 complex and the multi-subunit Ccr4/Not complex (Ccr4, Not1, Pop2, Caf40) (***Wolf and Passmore, 2014***). Specific mRNAs can recruit selected deadenylating factors after loss of the polyadenylate-binding protein (Pab1) that protects the mRNA from degradation on the 3′ end (***Finoux and Séraphin, 2006***; ***Goldstrohm et al., 2006***; ***Semotok et al., 2005***), but the choice of deadenylation pathway

eLife digest Cells contain a large group of DNA-like molecules called RNAs. While DNA stores and preserves information, RNA influences how cells use and regulate that information. As such, regulating the quantities of different RNAs is a key part of how cells survive, grow, adapt and respond to changes. For example, messenger RNAs (or mRNAs for short) carry genetic information from DNA which the cell reads to produce proteins. RNAs that are not needed can be degraded and removed from the cell by RNA degradation proteins.

Most RNA degradation proteins need to be able to bind to RNA in order to work. A technique called "photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation", often shortened to PAR-CLIP, can detect these proteins on their targets. The PAR-CLIP technique irreversibly links RNA-binding proteins to RNA and then collects those proteins and their bound RNAs for analysis. As with DNA, the RNAs can be identified using genetic sequencing. Degradation often starts at RNA ends, where specialized structures protect the RNA from accidental damage.

Using PAR-CLIP, Sohrabi-Jahromi, Hofmann et al performed a detailed study of 30 RNA degradation proteins in the yeast *Saccharomyces cerevisiae*. The results highlight the specialization of different proteins to different groups of RNAs. One group of proteins, for example, remove the protective 'cap' structure at the start of RNAs. Those mRNAs that are not efficiently producing proteins attracted a lot of these cap-removing proteins. The findings also identify proteins involved in RNA degradation in the cell nucleus – the compartment that houses most of the cell's DNA.

Together these findings provide an extensive data resource for cell biologists. It offers many links between different RNAs and their degradation proteins. Understanding these key cellular processes helps to reveal more about the mechanisms underlying all of biology. It can also shed light on what happens when these processes fail and the diseases that may result.

DOI: https://doi.org/10.7554/eLife.47040.002

remains unclear. Studies have pointed out a direct link between translation termination and mRNA degradation (reviewed in ***Huch and Nissan, 2014***), in particular deadenylation, which is dependent on Pab1 and Ccr4 (***Webster et al., 2018***). A proposed stepwise model for deadenylation suggests that first the average yeast polyA tail length of 90 nucleotides (nt) is reduced to 50 nt by the Pan2/Pan3 complex, before further shortening via the Ccr4/Not complex (***Beilharz and Preiss, 2007***; ***Brown and Sachs, 1998***; ***Tucker et al., 2001***). When the polyA tail reaches a length of 10–12 nt, the mRNA is decapped (***Chowdhury et al., 2007***; ***Tharun and Parker, 2001***), or subjected to exosome catalyzed degradation (***Bonneau et al., 2009***).

The second step in mRNA degradation is thought to be the removal of the 5′ cap by the decapping complex (Dcp1, Dcp2, Dcs1). The cap protects the mRNA from degradation by the 5′→3′ exonuclease Xrn1, which requires a 5′ monophosphate at the terminal residue (***Stevens and Poole, 1995***). Decapping is highly regulated by decapping enhancers such as the DEAD box helicase Dhh1, Edc2 and Edc3. Different potential mechanisms to trigger decapping include interference with translation initiation factors, facilitated assembly of the decapping machinery, and stimulation of Dcp2 catalytic function. Assembly of the decapping machinery occurs mainly after shortening of the polyA tail, which triggers decapping complex formation on the deadenylated 3′ end of mRNA and opening of the mRNA closed-loop structure (***Caponigro and Parker, 1995***; ***Morrissey et al., 1999***). In this closed-loop model, the 5′ and 3′ ends of the mRNA are thought to be in close proximity by forming a complex between translation initiation factors binding to the 5′ cap and Pab1 associated with the 3′ end, thereby contributing to mRNA expression regulation (***Vicens et al., 2018***; ***Wells et al., 1998***).

An alternative pathway of mRNA degradation after deadenylation is 3′→5′ degradation by the exosome and its auxiliary factors (***Anderson and Parker, 1998***). The exosome is a multi-subunit complex that consists of 10 core factors, comprising six members of the RNase PH protein family (Rrp43, Rrp45, Rrp42, Mtr3, Rrp41, Rrp46), three small RNA-binding proteins (Csl4, Rrp40, Rrp4) (***Allmang et al., 1999***), and the Rrp44/Dis3 protein, which harbors an exonuclease and an endonuclease domain (***Lebreton and Séraphin, 2008***; ***Schaeffer et al., 2009***). In addition to its functions in the cytoplasm, the exosome fulfills multiple roles in nuclear RNA processing and degradation

(*Lykke-Andersen et al., 2009*; *Ogami et al., 2018*) for which it is additionally bound by Rrp6, another 3′→5′ exonuclease, Rrp47, and Mpp6 (*Milligan et al., 2008*; *Mitchell et al., 2003*; *Synowsky et al., 2009*).

For RNA degradation by the exosome, RNA first passes through either the TRAMP or the Ski complex. TRAMP is a nuclear poly-adenylation complex (*Houseley and Tollervey, 2009*) that is involved in many of the RNA maturation and degradation processes and exists in two isoforms, TRAMP4 (Trf4, Air2 and Mtr4) and TRAMP5 (Trf5, Air1 and Mtr4). These complexes harbor a pA polymerase (Trf4 or Trf5), a zinc-knuckle putative RNA-binding protein (Air1 or Air2), and an RNA helicase (Mtr4). Defective nuclear RNAs are tagged with a short polyA tail by TRAMP, making them a more favorable substrate for the exosome core (*Vanácová et al., 2005*). The Ski complex is required for cytoplasmic exosomal degradation. The Ski7 protein is stably bound to the cytoplasmic exosome through the Ski4 subunit (*van Hoof et al., 2002*). The Ski2, Ski3, and Ski8 proteins form a subcomplex interacting with Ski7, which is required for 3′→5′ degradation of mRNAs (*Araki et al., 2001*; *Brown et al., 2000*; *Wang et al., 2005*). The Ski2 protein is an ATPase of the RNA helicase family that generates energy by ATP hydrolysis to unwind secondary structures and dissociate bound proteins to deliver the RNA to the exosome.

To degrade eukaryotic mRNAs that are defective in translation, cytoplasmic quality control mechanisms exist (*Doma and Parker, 2007*). Normal and aberrant mRNAs can be discriminated by the translation machinery, and translationally defective mRNAs are guided to a degradation pathway. mRNAs with aberrant translation termination due to a premature translation termination codon are subjected to nonsense-mediated decay (NMD) (*Losson and Lacroute, 1979*). Substrates for NMD are identified by the Upf1 protein interacting with the translation termination complex followed by binding of the Upf2 and Upf3 proteins, which enhances the helicase activity of Upf1 (*Baker and Parker, 2004*; *Chakrabarti et al., 2011*). During NMD, the mRNA can be subjected to enhanced deadenylation, deadenylation-independent decapping and rapid 3′→5′ degradation (*Cao and Parker, 2003*; *Mitchell and Tollervey, 2003*; *Muhlrad and Parker, 1994*).

The large variety of different RNA degradation factors poses the question how RNA degradation pathways are selected and whether the RNA sequence can influence this selection. Answering this question requires a systematic analysis of the RNA-binding profiles of the involved protein factors. Although several transcriptome profiles of the RNA degradation factors Xrn1, Rrp44, Csl4, Rrp41, Rrp6, Mtr4, Trf4, Air2 and Ski2 have been reported (*Delan-Forino et al., 2017*; *Milligan et al., 2016*; *Schneider et al., 2012*; *Tuck and Tollervey, 2013*), we lack transcriptome-wide binding profiles for components of the deadenylation, decapping, and NMD machineries, as well as other subunits of the exosome complex. Thus, the task of systematically analyzing the binding of subunits from all known factors involved in RNA degradation to a eukaryotic transcriptome ('transcriptome mapping') has not been accomplished thus far.

Here we used photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) to systematically generate transcriptome-wide protein binding profiles for 30 general RNA degradation factors in the yeast *Saccharomyces cerevisiae (S. cerevisiae)*. In-depth bioinformatic analysis and comparisons with previously reported PAR-CLIP data provide factor enrichment on different RNA classes and the binding behavior for mRNAs and their associated antisense transcripts. The results also give insights into how the various degradation complexes, and also different subunits in these complexes, may be involved in the degradation of different RNA species. Several conclusions can be drawn with respect to degradation pathway selection, new functions for known factors can be proposed, and several hypotheses emerge that can be tested in the future. Finally, our dataset provides a rich resource for future studies of eukaryotic RNA degradation pathways, mechanisms, and the integration of mRNA metabolism.

Results

Transcriptome maps for 30 RNA degradation factors

In order to get a better understanding of RNA processing and degradation in a eukaryotic cell, we measured transcriptome-wide binding locations of 30 RNA degradation factors involved in mRNA deadenylation, decapping, exosome-mediated degradation, and in RNA surveillance pathways including nuclear RNA surveillance and cytoplasmic nonsense-mediated decay (NMD) (*Figure 1A,B*).

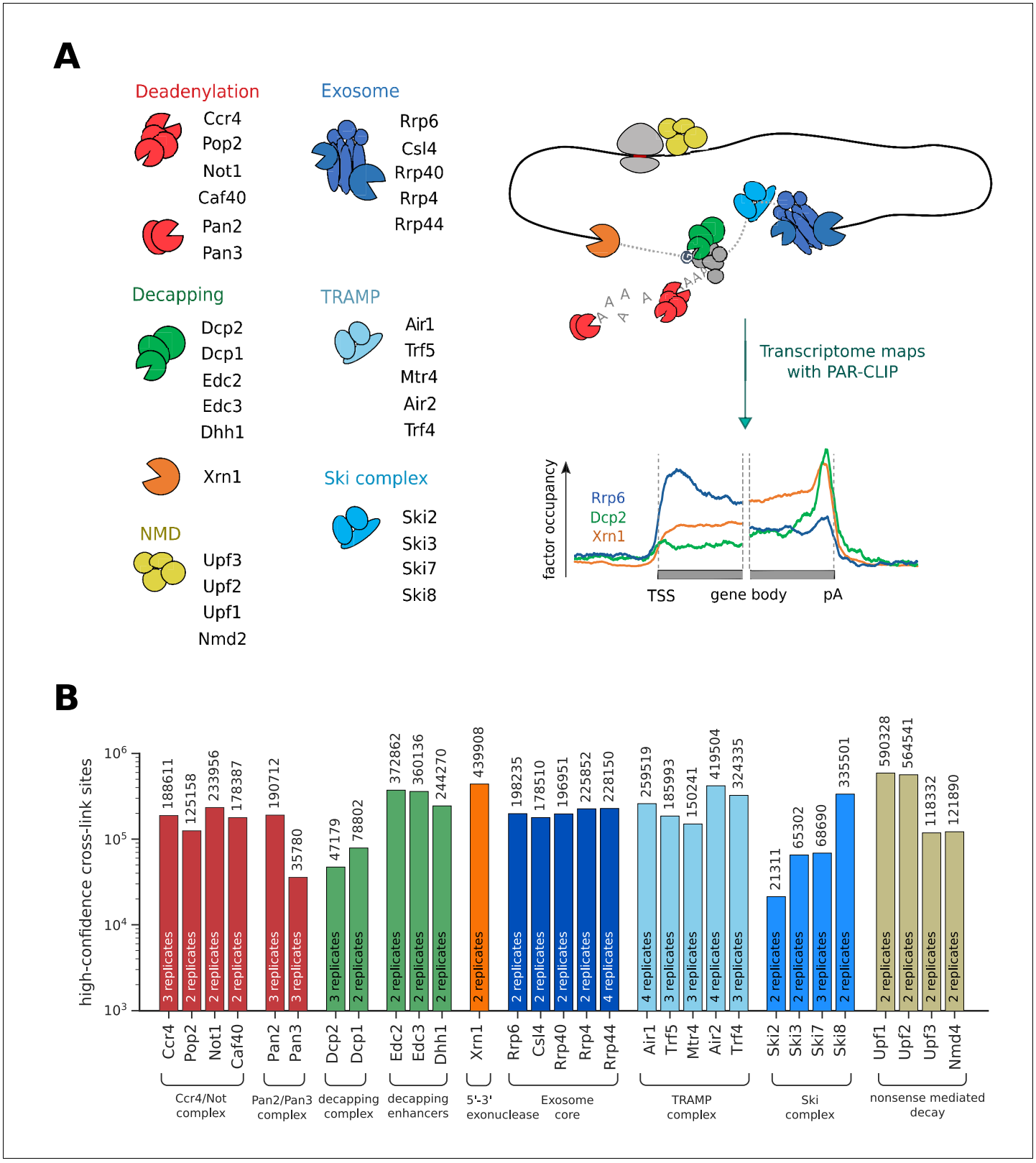


Figure 1. Overview of PAR-CLIP experiments performed in this study. (**A**) Overview of degradation pathways studied. (**B**) Number of high-confidence PAR-CLIP cross-link sites obtained for each factor after merging data from replicates.

DOI: https://doi.org/10.7554/eLife.47040.003

The following figure supplements are available for figure 1:

Figure 1 continued on next page

Figure 1 continued

Figure supplement 1. Biological replicate PAR-CLIP experiments have high correlation.
DOI: https://doi.org/10.7554/eLife.47040.004
Figure supplement 2. Western Blot analysis for all degradation factors analyzed in this study show IP efficiency.
DOI: https://doi.org/10.7554/eLife.47040.005

We performed PAR-CLIP in *S. cerevisiae* using our published protocol (***Battaglia et al., 2017***), with minor modifications (Materials and methods). The high reproducibility of these PAR-CLIP experiments is revealed by a comparison of two independent biological replicates that we collected for all 30 degradation factors (***Figure 1—figure supplement 1***), with Spearman correlations between 0.87 and 1.00 (mean: 0.94). We typically obtained tens of thousands of verified factor-RNA cross-link sites with p-values$\leq$0.005 (***Figure 1B***). These transcriptome maps represent an extensive, high-confidence dataset of in vivo RNA-binding sites for factors involved in RNA degradation.

Degradation factors exhibit transcript class specificity

We first compared degradation factor binding over different RNA classes. These included messenger RNA (mRNA), where we distinguished the 5′ untranslated region (5′ UTR), the coding sequence (CDS), introns, and the 3′ untranslated region (3′ UTR). We also included several classes of ncRNAs: ribosomal (r), transfer (t), small nucleolar (sno), and small nuclear (sn) RNAs, as well as stable unannotated transcripts (SUTs), cryptic unstable transcripts (CUTs), and Nrd1-unterminated transcripts (NUTs) (***Neil et al., 2009***; ***Pelechano et al., 2013***; ***Schulz et al., 2013***) (***Figure 2***).

A first analysis revealed that most PAR-CLIP sequencing reads fall into the mRNA transcript class, although many of the factors also show a considerable number of sequencing reads in ncRNAs, in particular rRNAs (***Figure 2A***). To obtain a more quantitative comparison, we defined log enrichment scores that reflect the preferences of factors in binding to a specific transcript class in comparison to other factors and classes. To correct for the different sizes of classes and different numbers of measured factor binding sites, we normalized the log enrichment scores by subtracting class- and factor-specific offsets, such that the mean for each class and each factor vanishes (***Figure 2B***, Materials and methods). This analysis highlights differences between degradation factors with respect to binding to various transcript classes, as will be discussed in detail below.

RNA end-processing complexes differ in their targets

The catalytic subunit Pop2 and the core subunits Not1 and Caf40 of the deadenylase complex Ccr4/Not have similar binding preferences for the 5′ UTR, the CDS and 3′ UTR of mRNAs, for rRNAs, tRNAs, snoRNAs, and snRNAs (***Figure 2B***, highlighted in red). Compared to other deadenylation factors of the Ccr4/Not complex, the catalytically active subunit Ccr4 has different binding preferences, and is strongly enriched at mRNA introns. The second deadenylation complex, Pan2/Pan3, shows a similar binding preference as the Ccr4/Not complex (except for the Ccr4 subunit), consistent with its dominant role in yeast mRNA deadenylation (***Boeck et al., 1996***). Pan3 shows a strong binding preference for rRNAs and tRNAs.

For all decapping-related factors we observed similar binding preferences among each other (***Figure 2B***, highlighted in green). They show the strongest enrichment at SUTs and at mRNAs compared to the other transcript classes. Decapping factors bind preferentially to CDS and 3′ UTR as well as SUTs. This is consistent with previous findings that SUTs are degraded via Dcp2-dependent pathways in the cytoplasm (***Marquardt et al., 2011***; ***Smith et al., 2014***; ***Thompson and Parker, 2007***). Dcp2, which harbors the hydrolase activity that removes the 5′ cap, and the decapping activator Edc3, additionally bind to NUTs. The 5′ exonuclease Xrn1 shows a similar binding preference as the decapping factors (***Figure 2B***, highlighted in orange). Taken together, complexes and enzymes that are known to target mRNA ends for 3′ deadenylation and 5′ decapping and degradation show remarkably distinct binding specificities to different transcript classes.

The exosome and surveillance factors

For the exosome we also observed binding to different RNA classes (***Figure 2B***, highlighted in royal blue). The core exosome subunits Csl4 and Rrp40 show similar cross-linking to rRNAs, tRNAs,

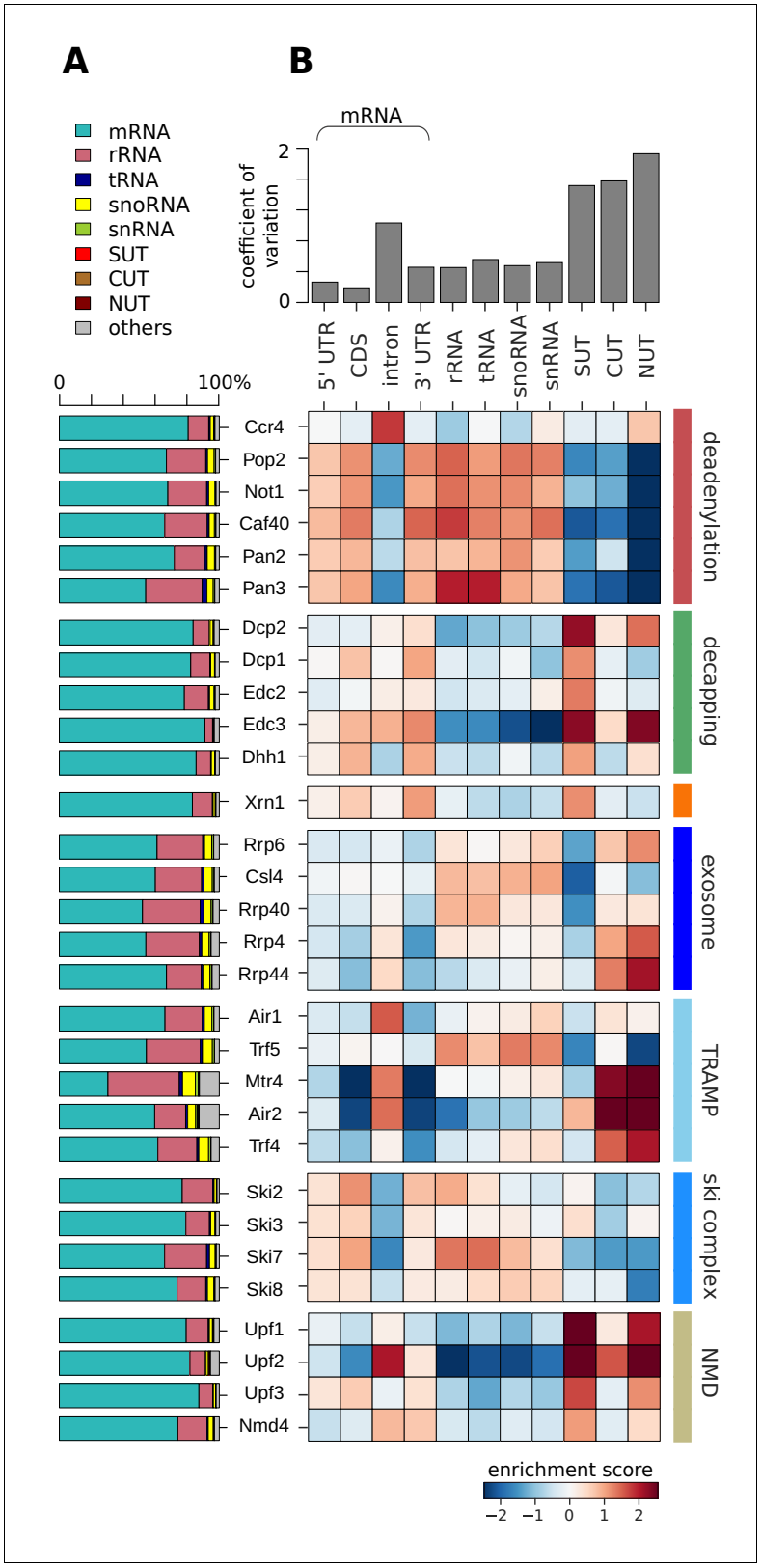


Figure 2. Distribution of degradation factor cross-link sites over the yeast transcriptome. (**A**) Fractions of high confidence PAR-CLIP sequencing reads of 30 yeast degradation factors fall into various transcript classes. Depicted classes are the following: messenger RNA (mRNA) in turquoise (n = 4,928), ribosomal RNA (rRNA) in antique pink (n = 24), transfer RNA (tRNA) in dark blue (n = 299), small nucleolar RNA (snoRNA) in yellow (n = 77),

Figure 2 continued on next page

Figure 2 continued

small nuclear RNA (snRNA) in green (n = 6), stable unannotated transcripts (SUTs) in red (n = 318), cryptic unstable transcripts (CUTs) in light brown (n = 637), Nrd1-unterminated transcripts (NUTs) in dark brown (n = 298)(Materials and methods). (**B**) Enrichment *z*-scores of high confidence PAR-CLIP cross-link sites of 30 yeast degradation factors (rows) in various segments of mRNA transcripts (left columns; UTR: untranslated region; intron; CDS: coding sequence), or other transcript classes as in A (other columns). The color-coded enrichment score shows the column and row normalized enrichment values of binding preferences of each factor for each transcript class (color encoded, depleted in blue and enriched in red). The coefficient of variation on top is the standard deviation divided by the mean for each transcript class.

DOI: https://doi.org/10.7554/eLife.47040.006

The following figure supplements are available for figure 2:

Figure supplement 1. Metagene profiles for subunits of the TRAMP complexes on snoRNA genes.

DOI: https://doi.org/10.7554/eLife.47040.007

Figure supplement 2. Different transcript classes have comparable U-content.

DOI: https://doi.org/10.7554/eLife.47040.008

snoRNAs, and snRNAs. The catalytic exosome subunit Rrp44 and the core subunit Rrp4 binds to introns of mRNAs, but preferentially to the short-lived, nuclear CUTs and NUTs. Rrp6, a subunit that is exclusively present in the nuclear exosome complex, shows binding to rRNAs, snoRNAs, snRNAs, CUTs and NUTs. This is consistent with the suggestion that the factor is needed for nuclear processing of such non-coding transcripts and degradation of short-lived nuclear transcripts (***Heo et al., 2013***; ***Vasiljeva and Buratowski, 2006***). This complex distribution of cross-links for different exosome subunits to different RNA classes reflects the distinct functions of the exosome in nuclear RNA surveillance, processing of stable ncRNAs, and cytoplasmic mRNA degradation (***Zinder and Lima, 2017***).

The two TRAMP complexes TRAMP4 and TRAMP5 show clearly distinct cross-linking patterns (***Figure 2B***, highlighted in light blue). TRAMP4 subunits (Mtr4, Air2, Trf4) are enriched in introns, consistent with a function on mRNAs, and on SUTs, CUTs, and NUTs. The TRAMP5 complex (Mtr4, Air1, Trf5) shows binding enrichment for introns, rRNAs, tRNAs, snRNAs, and snoRNAs. This is in agreement with previous data, which showed rRNA binding for Mtr4 and exosome subunits (***Delan-Forino et al., 2017***; ***Schneider and Tollervey, 2013***). Moreover, the TRAMP complex cooperates with the Nrd1/Nab3 complex and the nuclear exosome complex during the maturation and 3´ pre-processing of snoRNAs (***Grzechnik and Kufel, 2008***). To distinguish binding upon degradation and binding in order to pre-process snoRNAs, we investigated metagene profiles of TRAMP subunits along snoRNA genes (***Figure 2—figure supplement 1***). Air1/Trf5 bind almost exclusively to the gene body whereas Air2/Trf4 bind downstream of the 3´ end. This suggests that TRAMP5 is mainly involved in snoRNA degradation, whereas TRAMP4 may work together with the Nrd1/Nab3 machinery to pre-process snoRNAs (***Figure 2—figure supplement 1***) and to target NUTs, SUTs, and CUTs for degradation (***Figure 2B***, highlighted in light blue).

The cross-linking preferences of subunits of the Ski complex differ only slightly from each other (***Figure 2B***, highlighted in cyan). All Ski complex subunits bind the 5´ UTR, CDS, and 3´ UTR of mRNAs, rRNAs, tRNAs, snoRNAs, and snRNAs. The Ski2 subunit preferentially binds to the CDS of mRNAs, consistent with its function as a helicase to detach bound proteins from the mRNAs (***Houseley and Tollervey, 2009***; ***Lebreton and Séraphin, 2008***). The exosome adaptor subunit Ski7 preferentially binds rRNAs and tRNAs. These patterns are consistent with the model that the exosome cooperates with distinct accessory complexes and factors to target different transcript classes. Finally, we observed similar cross-linking patterns for all NMD factors with strong binding to SUTs and NUTs (***Figure 2B***, highlighted in yellow). Upf2 shows an additional binding preference to introns and CUTs. Upf3 also binds to the 5´ UTR, CDS, and 3´ UTR of mRNAs, and Nmd4 binds to introns and 3´ UTRs of mRNAs.

Distinct factor distribution along mRNA

We next focused on degradation factor distribution on mRNAs. We prepared metagene profiles showing the average occupancy of each factor around the mRNA transcription start sites (TSS) and the poly-adenylation (pA) sites, respectively (***Figure 3***). The Pan2/Pan3 deadenylase complex and

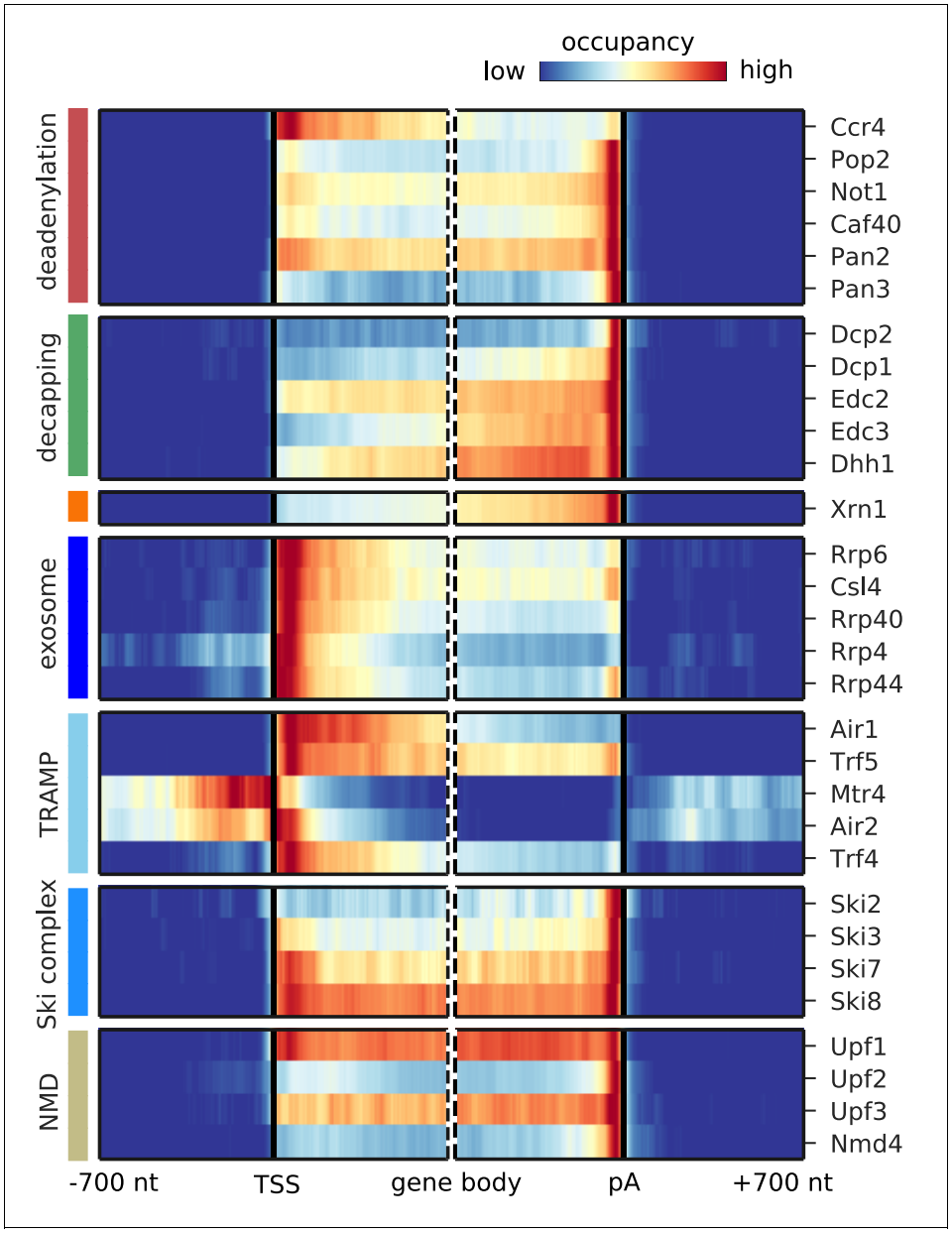


Figure 3. Metagene analysis of degradation factor binding on mRNAs. Averaged occupancy profiles of degradation factors over mRNAs aligned around their transcription start site (TSS) (n = 3,193, left) and around their poly-adenylation (pA) site (n = 3,193, right) in a window of [±700 nt]. Regions that have neighboring transcripts on the same strand were removed to avoid contaminating profiles (Materials and methods). Factors are grouped according to their functional role; from top to bottom: deadenylation, decapping, Xrn1, exosome, TRAMP complex, Ski complex, or NMD. The color code shows the average occupancy normalized between the minimum (blue) and maximum (red) values per profile.

DOI: https://doi.org/10.7554/eLife.47040.009

The following figure supplements are available for figure 3:

Figure supplement 1. Metagene profiles of yeast RNA degradation factors centered on translation start and stop sites in comparison to TIF-annotated TSS and pA sites.

DOI: https://doi.org/10.7554/eLife.47040.010

Figure supplement 2. Comparison of binding profiles on genes containing annotated upstream sense NUTs with all mRNAs.

DOI: https://doi.org/10.7554/eLife.47040.011

Figure 3 continued on next page

Figure 3 continued

Figure supplement 3. Metagene analysis of degradation factor binding on mRNAs after removing signals from known NUTs and CUTs.

DOI: https://doi.org/10.7554/eLife.47040.012

the Ccr4/Not subunits Pop2, Not1, and Caf40 all cross-link upstream of the 3´ end of mRNA with the highest enrichment at the pA site, as expected from their function in shortening the polyA tail. The catalytic subunit Ccr4 binds strongly in the 5´ region of mRNAs. All 5´ decapping factors bind upstream of the pA site, and all but the catalytically active subunit Dcp2 show increasing occupancy towards the 3´ end of mRNAs. These patterns can be explained if decapping factors are pre-bound to mRNAs that form a closed loop that holds the RNA ends in proximity. In contrast, Dcp2 binds almost exclusively at the pA site, suggesting that it might be recruited only upon active mRNA degradation. The cytoplasmic 5´ exonuclease Xrn1 has the highest occupancy towards the 3´ end, similar to the previously published crosslinking and cDNA analysis (CRAC) data (***Tuck and Tollervey, 2013***), thereby resembling the binding profiles of the decapping factors. Comparison of the binding profiles aligned at the pA site or alternatively with profiles aligned at the translation stop codon shows that the binding preference indeed lies at the end of the 3´ UTR independent of the stop codon position (***Figure 3—figure supplement 1B,C***).

The exosome core subunits (Csl4, Rrp40, and Rrp4) and the catalytically active subunits (nuclear: Rrp6, cytoplasmic: Rrp44) cross-link to the 5´ end of the transcript (***Figure 3***), possibly because the exosome binds to the 5´ end while digesting the 3´ end, or more likely because the exosome slows down towards the remaining 5´ end of mRNAs after rapid degradation from the 3´ end. Both TRAMP complexes bind mainly in the 5´ region of mRNAs near the TSS, as previously observed for Mtr4 and Trf4 (***Tuck and Tollervey, 2013***).

The Ski complex components Ski7 and Ski8 occupy the entire mRNA with increasing occupancy towards the pA site, whereas Ski2 and Ski3 show more discrete binding towards the polyA tail (***Figure 3***). The NMD factors Upf1 and Upf3 show binding over the entire mRNA with highest occupancy at the pA site, consistent with their role in scanning for premature stop codons in mRNAs and remodeling of the 3´ end of protein-RNA complexes and completion of mRNA decay (***Franks et al., 2010***). In addition, Upf2 and Nmd4 show strongest binding near the 3´ ends of mRNAs. Taken together, the distribution of cross-links along mRNA transcripts differs between degradation complexes and in some cases also between their subunits.

Surveillance of aberrant nuclear ncRNA

Pervasive transcription of the genome leads to many short-lived aberrant RNAs that must be rapidly detected and degraded in the nucleus. We previously reported that the RNA surveillance factors Nrd1 and Nab3 strongly cross-link to aberrant upstream antisense RNA that stems from bidirectional transcription (***Schulz et al., 2013***). In order to find factors cross-linking to aberrant ncRNAs, we plotted the occupancy of all 30 investigated factors on the antisense strand of known mRNAs (***Figure 4***). For comparison, we plotted the published Nrd1 and Nab3 profiles in the first two lanes of ***Figure 4***. The factors are involved in processing and degradation of Nrd1-unterminated transcripts, or NUTs (***Schulz et al., 2013***), and are expected to show similar binding to upstream antisense RNA as Nrd1 and Nab3. Indeed, we observed a similar binding pattern for all exosome subunits (Rrp6, Csl4, Rrp40, Rrp4, Rrp44) and subunits of the TRAMP4 complex (Mtr4, Air2, Trf4). Consistent with this, these factors also bind strongly to previously annotated NUTs and CUTs (***Figure 2***) and show strong enrichment of Nrd1 and Nab3 motifs (GTAG, CTTG) around their cross-link sites (***Figure 4—figure supplement 1***).

It has been shown that Nrd1 is involved in terminating transcripts upstream of the TSS. We also observe a strong signal for binding upstream of TSS on the sense strand for Air2 and Mtr4 (***Figure 3***). This suggests that the TRAMP4 complex is involved in degradation of those Nrd1-regulated upstream sense transcripts. To investigate this hypothesis, we compared the binding profiles around the TSS of 459 protein-coding genes, previously annotated as having upstream Nrd1-unterminated transcripts, or NUTs (***Schulz et al., 2013***), with the profiles obtained for all mRNAs (***Figure 3—figure supplement 2A,B***). TRAMP4 and the exosome subunits show a strong preference for binding to the upstream promoter region of genes that are controlled by the Nrd1/Nab3 complex. To further

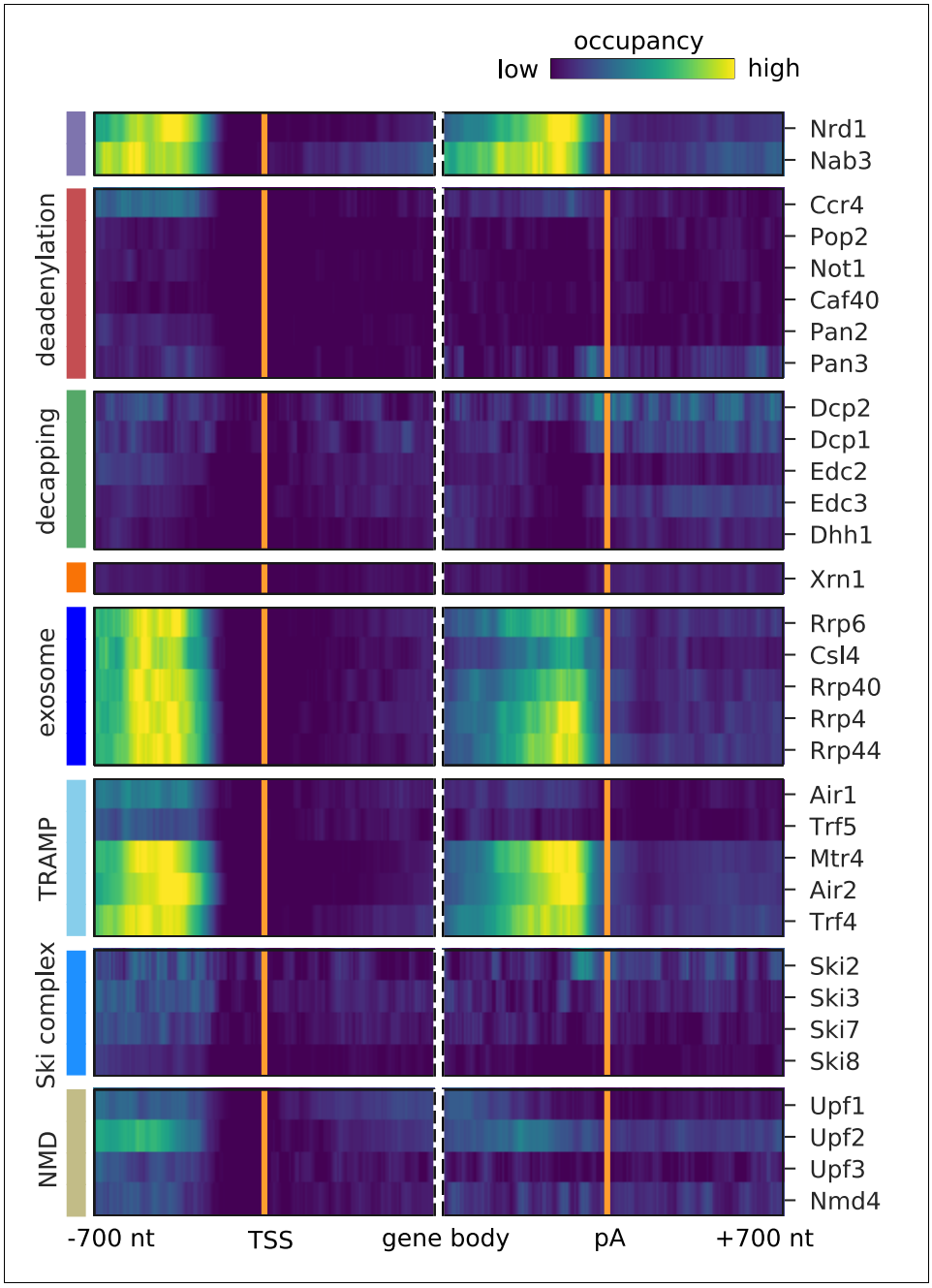


Figure 4. Surveillance of aberrant nuclear antisense RNAs by the exosome and the TRAMP4 complex. Averaged occupancy profiles of degradation factors binding to transcripts antisense of mRNAs aligned around transcription start site (TSS) (n = 3,076, left) and around their poly-adenylation (pA) site (n = 2,705, right) in a window of [±700 nt]. Regions with annotated genes on the antisense strand are removed to avoid contaminating the profiles (Materials and methods). The color code shows the average occupancy normalized between the minimum (blue) and maximum (yellow) values per profile. On top, previously published PAR-CLIP profiles for Nrd1 and Nab3 are included for comparison (***Schulz et al., 2013***).

DOI: https://doi.org/10.7554/eLife.47040.013

The following figure supplements are available for figure 4:

Figure supplement 1. Motif enrichment analysis shows enrichment of Nrd1/Nab3 motifs for the TRAMP4 and the exosome complex.

DOI: https://doi.org/10.7554/eLife.47040.014

Figure 4 continued on next page

Figure 4 continued

Figure supplement 2. The aberrant nuclear ncRNAs bound by components of the exosome and the TRAMP4 complex are primarily NUTs and CUTs.

DOI: https://doi.org/10.7554/eLife.47040.015

confirm that this upstream signal originates from NUTs and CUTs, we excluded cross-link sites that fall within such previously annotated regions. We then compared the binding profiles generated from the remaining binding sites on mRNAs (***Figure 3—figure supplement 3***). Upon filtering, the signal upstream of the TSS for Air2 and Mtr4 decreases, showing that Nrd1-mediated regulation is the primary cause for this upstream signal. Comparison of the observed antisense profiles (***Figure 4***) with those obtained after excluding cross-link sites in previously annotated NUT and CUT regions (***Figure 4—figure supplement 2***) confirms that most of the signal originates from transcripts that are targeted by the Nrd1/Nab3 machinery.

These results are consistent with the idea that the nuclear RNA surveillance machinery involves, in addition to Nrd1 and Nab3, the TRAMP4 complex and the nuclear exosome. Indeed, it was reported that TRAMP4 can add a short polyA tail on aberrant RNAs (***Wyers et al., 2005***), which may trigger degradation by the nuclear exosome. It was also recently shown that Nrd1 and Trf4 interact, providing a basis for coupling surveillance-mediated termination to RNA degradation (***Tudek et al., 2014***).

Interactions between RNA processing machineries

To find out which groups of factors can work together in degrading transcripts, we analyzed their tendency to co-occupy the same transcripts by calculating the Pearson correlation of their occupancy across all transcripts (***Figure 5A***). We also analyzed their co-localization, that is the tendency of a factor to bind near to another factor's binding sites using a range of ±40 nt from each cross-link site (***Figure 5B***). To relate these profiles to those of other factors, we included previously published PAR-CLIP profiles from our lab (***Supplementary file 1***). Profiles were available for factors that function in nuclear RNA surveillance (Nrd1, Nab3), cap binding (Cbc2), mRNA transcript elongation (Bur1, Bur2, Ctk1, Ctk2, Cdc73, Ctr9, Leo1, Paf1, Rtf1, Set1, Set2, Dot1, Spt5, Spt6, Rpb1), pre-mRNA splicing (Ist3, Nam8, Mud1, Snp1, Luc7, Mud2, Msl5), pre-mRNA 3´ processing (Pab1, Pub1, Rna15, Mpe1, Cft2; Yth1), transcription termination (Rat1, Rai1, Rtt103, Pcf11), and mRNA export (Hrp1, Tho2, Gbp2, Hrb1, Mex67, Sub2, Yra1, Nab2, Npl3) (***Baejen et al., 2017***; ***Baejen et al., 2014***; ***Battaglia et al., 2017***; ***Schulz et al., 2013***).

Co-occupancy and co-localization plots for all factors can be found in ***Figure 5—figure supplements 1*** and ***2***, respectively. A two-dimensional embedding of co-occupancy profiles between all these processing factors is shown in ***Figure 5C***. It represents the degree of similarities between co-occupancy of transcripts (***Figure 5A***) in terms of the distance in two dimensions. The two-dimensional embedding of the co-localization matrix in ***Figure 5B*** shows a similar clustering (***Figure 5—figure supplement 3***). This extensive global analysis suggests which factors reside in functional complexes and which functional complexes may interact during RNA processing and degradation. The analysis recovers several established interactions between subunits of known complexes and between different complexes, providing a positive control. For example, all factors of the decapping complex show very high co-occupancy and co-localization, as do Air2 and Mtr4, which reside in the TRAMP4 complex.

The analysis contains a lot of new information, forcing us to focus here on a few interesting, novel findings (***Figure 5C***). First, the largest cluster is formed by the previously analyzed factors involved in transcription elongation by RNA polymerase II (cluster 1) and in co-transcriptional pre-mRNA processing, including cap-binding complex (Cbc2), 3´ processing, transcription termination, and RNA export. The degradation factors Ccr4 and Air1 also reside in this cluster, maybe reflecting the role of Ccr4 in transcription elongation (***Kruk et al., 2011***). A second cluster is formed by splicing factors (cluster 2). Factors involved in nuclear and cytoplasmic exosomal degradation (Rrp6, Csl4, Rrp4, Rrp40 and Rrp44) form a third cluster (cluster 3). Close to cluster 3, we find the TRAMP4 complex subunit Trf4, the elongation factors Dot1, Paf1, Leo1, and the termination factors Pcf11 and Rai1. Rai1 has been shown to detect and remove incomplete 5´ cap structures, to subject aberrant pre-mRNAs to nuclear degradation (***Jiao et al., 2010***).

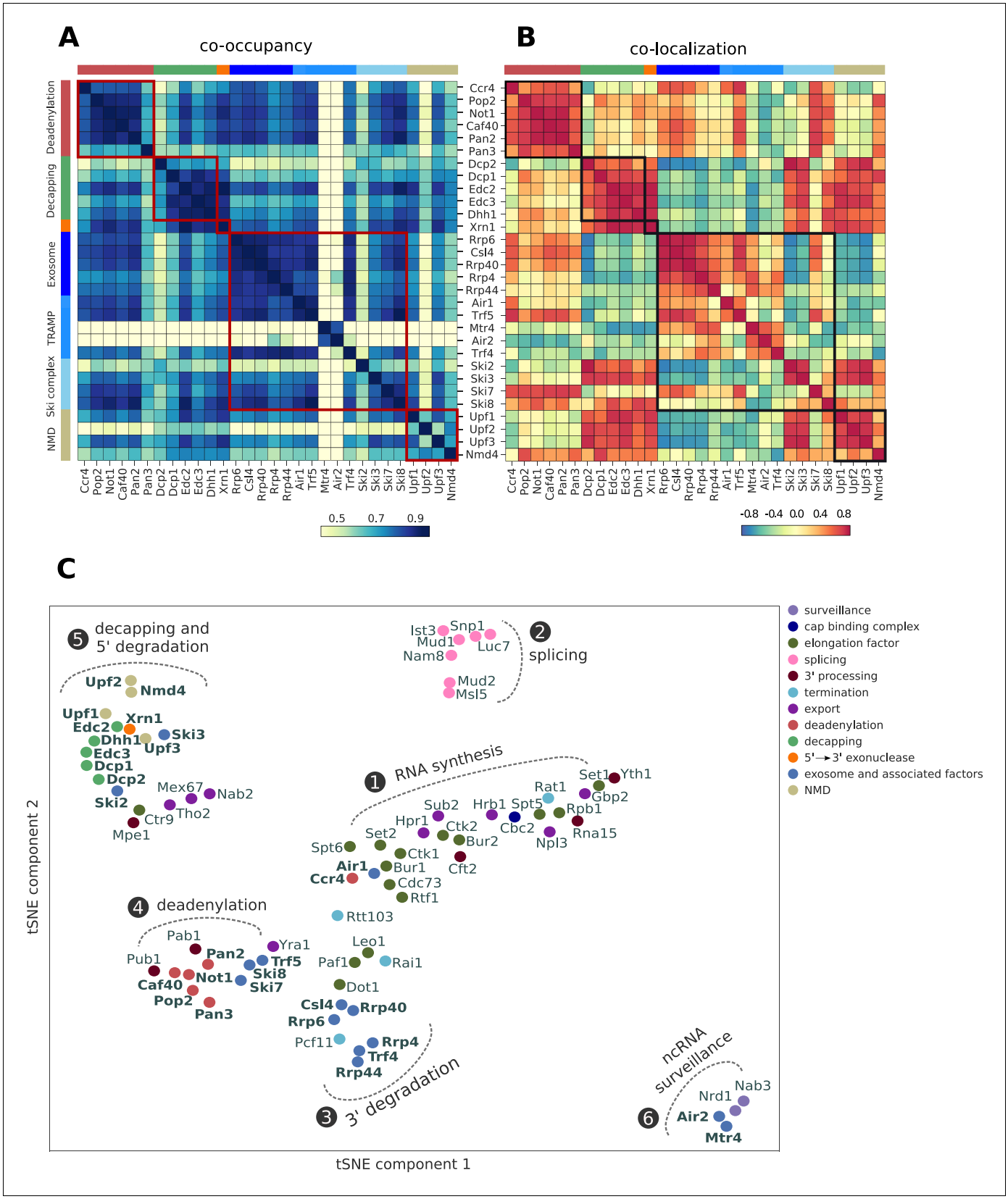


Figure 5. Global co-occupancy and co-localization analysis reveals unexpected cooperation between factors from different complexes and pathways. (**A**) Matrix of pairwise correlation coefficients of factor occupancies evaluated over all transcripts. (**B**) Matrix of co-localization based on the enrichment of factor x binding within 40 nt of the cross-link site of factor x′ (Materials and methods). (**C**) Two-dimensional embedding of the co-occupancies in (**A**)

Figure 5 continued on next page

Figure 5 continued

analyzed for 74 RNA processing factors with tSNE, including 30 factors from this study (highlighted in bold), and 44 factors from previous studies (***Baejen et al., 2017***; ***Baejen et al., 2014***; ***Battaglia et al., 2017***; ***Schulz et al., 2013***) (***Supplementary file 1***). Factors that are plotted in close proximity show a preference for binding to the same transcripts. Clusters present factors involved in RNA synthesis (1), splicing (2), 3´ processing (3), deadenylation (4), decapping (5), nuclear ncRNA processing (6), and surveillance (7).

DOI: https://doi.org/10.7554/eLife.47040.016

The following figure supplements are available for figure 5:

Figure supplement 1. Co-occupancy for 74 RNA processing factors.

DOI: https://doi.org/10.7554/eLife.47040.017

Figure supplement 2. Co-localization coefficients for all 74 RNA processing factors.

DOI: https://doi.org/10.7554/eLife.47040.018

Figure supplement 3. Two-dimensional embedding of co-localization between 74 RNA processing factors.

DOI: https://doi.org/10.7554/eLife.47040.019

A fourth cluster is formed by mRNA deadenylation factors together with polyA tail binding proteins (Pab1 and Pub1), Ski7, Ski8, Trf5, and the export factor Yra1 (cluster 4). This is consistent with coupled mRNA deadenylation and subsequent degradation from its 3´ end by the exosome with the Ski or TRAMP complex as adaptors. The fifth cluster is formed by mRNA decapping factors, which cluster together with Xrn1, suggesting a coupling of mRNA decapping with degradation from the 5´ end by Xrn1 (cluster 5). The NMD-involved factors Upf1, Upf2, Upf3 and Nmd4, and Ski2 and Ski3 are also found in cluster 5. The high correlation between Xrn1 and Ski2 has been reported in a CRAC experiment (***Tuck and Tollervey, 2013***). The elongation factor Ctr9, the 3´ processing factor Mpe1 and the export factors Tho2, Mex67 and Nab2 are also found in cluster 5. A last cluster (cluster 6) is formed by factors involved in nuclear RNA surveillance, including Air2, Mtr4 and the Nrd1/Nab3 complex. Taken together, these findings are consistent with known functional associations and physical interactions between factors and suggest intriguing new associations to be investigated in future work.

5´ degradation machinery senses translation efficiency

To study the link between cytosolic mRNA translation and degradation, we compared the occupancy of degradation factors on mRNAs to their average codon-optimality score ('transcript optimality') (***Figure 6A***, ***Figure 6—figure supplements 2–8A***). We found that the 5´ decapping machinery and Xrn1 preferentially bind transcripts with low transcript optimality. In contrast, the 3´ deadenylation machinery and the exosome bind more strongly to optimal transcripts. We asked whether this correlation with codon optimality is introduced by only a few differentially bound codons or by global enrichment/depletion of optimal codons. For this purpose, we introduced a 'codon enrichment score', which measures a codon's enrichment in the set of transcripts bound by the factor relative to the yeast mRNA pool. For Dcp2 this enrichment score is high on non-optimal codons, and low on optimal codons, whereas the opposite trend is observed for Ccr4 and most degradation factors (***Figure 6B***, ***Figure 6—figure supplement 1–7***) This is consistent with a model that ribosome stalling on translationally inefficient codons can lead to recruitment of Dcp2 and Xrn1 and subsequent 5´ degradation of the transcript (***Heck and Wilusz, 2018***).

To investigate the significance of the correlation between transcript optimality and binding of the 5´ degradation machinery, we compared the contribution of several mRNA features in explaining the occupancy patterns retrieved from PAR-CLIP experiments. Since mRNA expression, half-life, and translation optimality are inter-correlated (***Figure 6—figure supplement 8***), a causative effect of one of these features on binding strength may lead to correlations with all three features. To better distinguish correlation from causation, we used linear regression analysis to explore whether correlations between factor binding and optimality are better explained with other mRNA features (***Figure 6—figure supplement 9***). We assessed the significance of features via the likelihood ratio test on the multi-variate linear regression model for occupancy. The likelihood ratio test calculates the significance of a feature from the change of the likelihood (quantifying the prediction quality) upon removal of that feature from the regression model. For decapping enhancers (Edc2, Edc3, and Dhh1) and Xrn1, low codon optimality is the most determining feature for binding (***Figure 6C***). The same is true for NMD factors Upf1 and Upf3, which are known to bind non-optimal transcripts

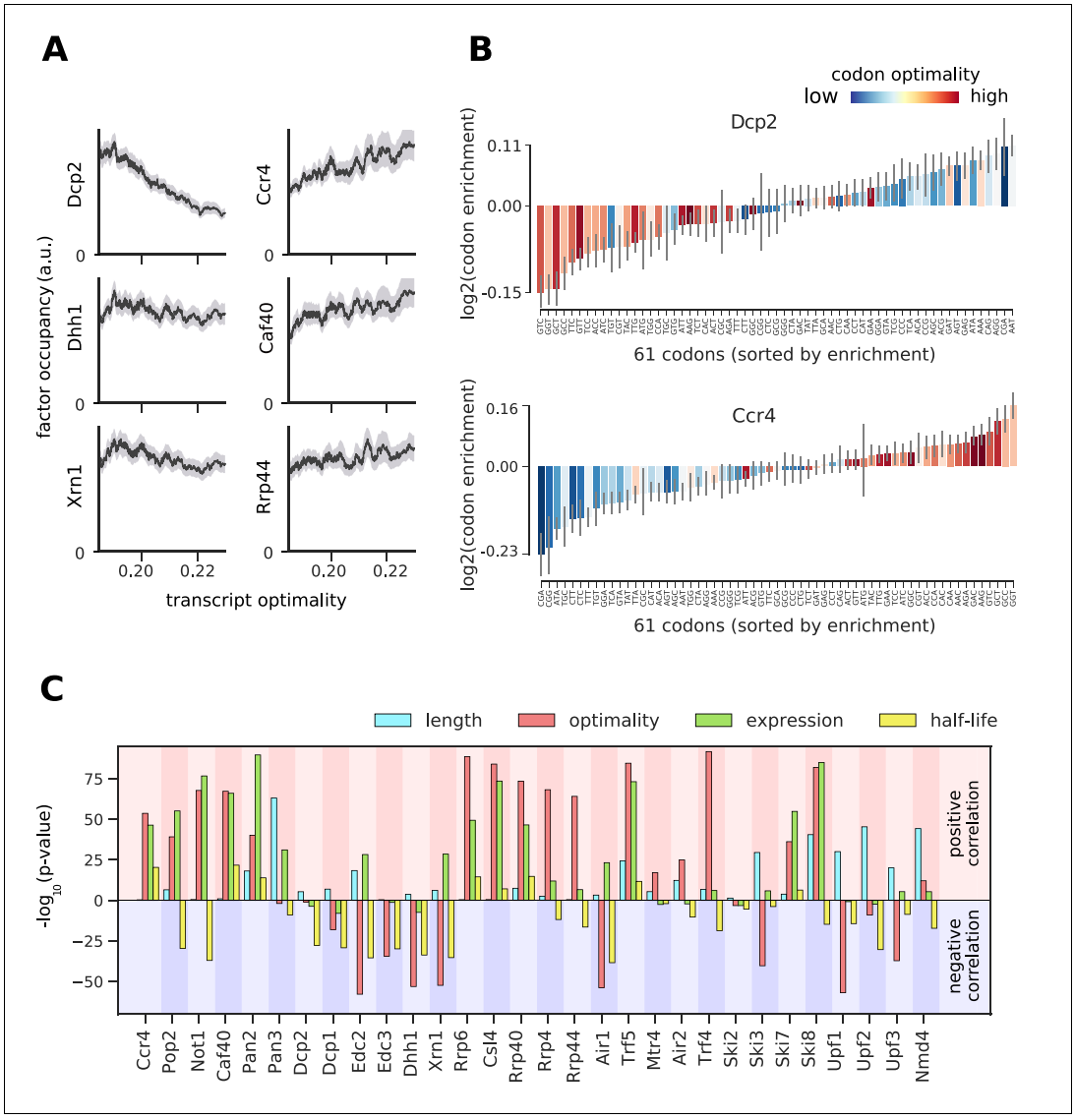


Figure 6. Binding preferences reveal a link between decapping-mediated degradation and translation. (**A**) Total occupancy per mRNA (according to TIF-seq annotation) for six factors as a function of the average mRNA codon optimality (transcript optimality). The occupancy of factors from the 5′→3′ degradation machinery (decapping and Xrn1, left) decreases with increasing transcript optimality, whereas the occupancy of factors from the 3′→5′ degradation machinery (Ccr4 and Caf40 deadenylation complex subunits and exosome subunit Rrp44, right) increases with increasing average codon optimality. (Gray shading: 95% confidence intervals generated by bootstrapping mRNAs). (**B**) Codon enrichment in transcripts bound by Dcp2 and Ccr4 compared to the average frequency over all mRNAs. The bar colors represent codon optimality, with highly optimal codons shown in dark red. (Thin gray lines: 90% confidence intervals generated by bootstrapping coding sequences.) (**C**) Significance of correlations between the binding strength of degradation factors and transcript length, transcript optimality (***Pechmann and Frydman, 2013***), expression level (***Baejen et al., 2017***), and half-life derived by multivariate linear regression analysis (Materials and methods). Bars are separated according to the direction of correlation with positive correlation marked by a red background and negative correlation marked by a blue background.

DOI: https://doi.org/10.7554/eLife.47040.020

The following figure supplements are available for figure 6:

Figure supplement 1. Occupancies of deadenylation factors (Ccr4, Pop2, Not1, Caf40, Pan2, and Pan3) compared to transcript length, optimality, expression level, and half-life.

DOI: https://doi.org/10.7554/eLife.47040.021

Figure supplement 2. Occupancies of decapping factors (Dcp2, Dcp1, Edc2, Edc3, and Dhh1) compared to transcript length, optimality, expression level, and half-life.

DOI: https://doi.org/10.7554/eLife.47040.022

Figure supplement 3. Occupancy of Xrn1 compared to transcript length, optimality, expression level, and half-life.

DOI: https://doi.org/10.7554/eLife.47040.023

Figure 6 continued on next page

Figure 6 continued

Figure supplement 4. Occupancies of exosome components (Rrp6, Csl4, Rrp40, Rrp4, and Rrp44) compared to transcript length, optimality, expression level, and half-life.

DOI: https://doi.org/10.7554/eLife.47040.024

Figure supplement 5. Occupancies for components of the TRAMP complex (Air1, Trf5, Mtr4, Air2, and Trf4) compared to transcript length, optimality, expression level, and half-life.

DOI: https://doi.org/10.7554/eLife.47040.025

Figure supplement 6. Occupancies for components of the Ski complex (Ski2, Ski3, Ski7, and Ski8) compared to transcript length, optimality, expression level, and half-life.

DOI: https://doi.org/10.7554/eLife.47040.026

Figure supplement 7. Occupancies for components of the NMD pathway (Upf1, Upf2, Upf3, and Nmd4) compared to transcript length, optimality, expression level, and half-life.

DOI: https://doi.org/10.7554/eLife.47040.027

Figure supplement 8. Distributions of transcript length, half-life, expression level and transcript optimality for yeast mRNAs.

DOI: https://doi.org/10.7554/eLife.47040.028

Figure supplement 9. Correlation between binding to degradation factors and transcript length, codon-optimality, expression, and half-life.

DOI: https://doi.org/10.7554/eLife.47040.029

(***Celik et al., 2017***). This result confirms the importance of the translation efficiency for the stability of cytosolic mRNAs and strengthens our finding that transcripts with low average codon optimality are preferentially targeted by the decapping machinery and degraded from the 5′ end.

Decapping factors are enriched upon RNA degradation

Although decapping occurs at the 5′ end of mRNAs, decapping factors show a strong occupancy near the 3′ end (***Figure 3***). To investigate this further, we compared metagene profiles of decapping factors between stable (top 25%) and unstable (bottom 25%) transcripts, using mRNA half-life estimates (***Figure 7A***, Materials and methods). On both stable and unstable mRNAs, Dcp1, Edc2, Edc3, and Dhh1 show increased binding near the 3′ end, but unstable RNAs show a higher occupancy in the transcript body. The catalytically active subunit Dcp2 binds almost exclusively at the 3′ end and has a higher occupancy on unstable transcripts. Moreover, A-rich 4-mers are abundant around the proximity (eight nt) of Dcp2-cross-link sites (***Figure 7C***), indicating a binding preference of Dcp2 for A-rich RNA sequences. Overall, these binding patterns suggest that decapping factors are bound in transcript bodies and near the 3′ end of transcripts, and that through closed-loop formation of the mRNA they are in close proximity to the 5′ end. Decapping factors might also travel with the 5′→3′ exonuclease Xrn1 upon RNA degradation.

Decapping factors may bind to complete mRNAs or to transcripts that are in the process of being degraded. To quantify these two behaviors, we combined our PAR-CLIP occupancy data with RNA half-life estimates (Materials and methods). We modeled the occupancy of factors on mRNA as the sum of binding to all transcripts (b) and surplus binding to transcripts that are in the process of degradation ($\frac{\mathrm{a}}{\mathrm{t}_{1/2}}$). Therefore, we can model occupancy as a function of half-life with a linear equation ($\mathrm{occupancy} = \frac{\mathrm{a}}{\mathrm{t}_{1/2}} + \mathrm{b}$). In cases where there is no surplus binding upon active degradation, that is the occupancy is the same as in intact RNAs, 'a' will be zero. For 5′ decapping factors, this model closely fits the occupancy patterns retrieved from our experiments (***Figure 7B***), other degradation factors also follow this pattern to varying degrees (***Figure 6—figure supplements 1–7***). In particular, Dcp2 shows a very high a/b ratio, revealing that it cross-links preferentially to transcripts that are being degraded. This analysis strongly suggests that the 5′ decapping machinery, although present to some extent on complete mRNAs, is enriched when mRNAs are degraded.

Discussion

Here we report transcriptome-wide binding maps for 30 RNA degradation factors in yeast. A detailed bioinformatics analysis of these maps revealed how degradation factors vary in their binding specificities for different classes of RNAs and with respect to their preferred locations on RNA transcripts. Global comparisons of the profiles alongside previously published profiles of other RNA-binding factors revealed clusters of factors that co-occupy RNAs or co-localize on RNAs. Our data

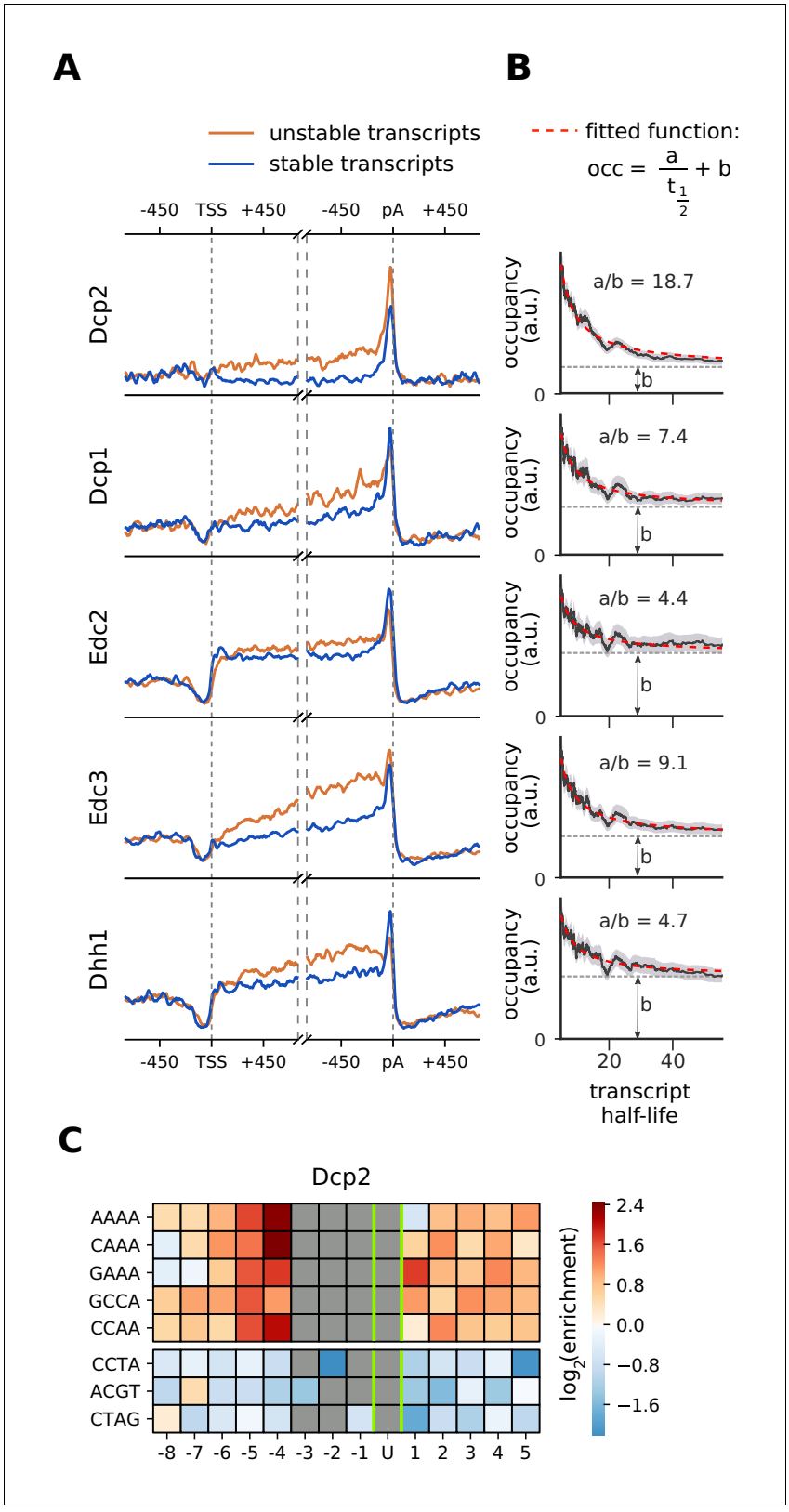


Figure 7. Location and recruitment of the decapping complex Dcp1/Dcp2 and decapping enhancers Edc3, Dhh1, and Edc2. (**A**) Smoothed, transcript-averaged PAR-CLIP occupancy profiles aligned at TSS and pA sites [±750 nt] of unstable and stable transcripts (first and fourth quantile of half-life distribution, respectively). (**B**) Dependence of total occupancy of factors on the transcripts half-life. The fitting function is plotted in red and the fitted value for b

Figure 7 continued on next page

Figure 7 continued

is marked with a dashed gray line. (Gray shade: 95% confidence intervals generated by bootstrapping transcripts). (**C**) Sequence binding preference for the catalytically active subunit of decapping complex (Dcp2), illustrated with the five most enriched and the 3 most depleted 4-mers. The color code shows the log2 enrichment factor of 4-mers around PAR-CLIP cross-link sites [±8 nt]. Dark red represents strong enrichment and dark blue shows strong depletion of a 4-mer. Infeasible combinations are shown with gray. The most highly enriched field is binding AAAAU with the cross-link at the U, which is enriched over random expectation approximately $2^{2.3}$ = 5-fold.

DOI: https://doi.org/10.7554/eLife.47040.030

are consistent with a large body of published results and extend these to a global scale. In addition, the results revealed several unexpected, novel findings, which we discuss below. Although our data reflect factor cross-linking signal and measure occupancy on transcripts, and do not directly reveal function, the correlations of occupancies between factors and with transcript properties indicate functional aspects and suggest functional associations between factors that may guide future studies.

With respect to canonical mRNA turnover in the cytoplasm, the Pan2/Pan3 deadenylation complex and subunits of the Ccr4/Not complex bound preferentially to the pA site, reflecting a function in polyA tail shortening at early stages of RNA degradation. In contrast, the Ccr4/Not complex subunit Ccr4 bound also strongly in the 5′ region of transcripts. This pattern may reflect functional differences between Ccr4 and Pop2 during deadenylation (***Webster et al., 2018***) or an additional function of Ccr4 in transcription elongation (***Kruk et al., 2011***). Ccr4/Not subunits also show variations in their RNA-binding specificities, suggesting several isoforms of the complex that vary in composition and function, or RNA-specific conformational rearrangements. Factors involved in decapping show higher cross-linking near the RNA 3′ end, consistent with previously proposed binding near the pA site (***Chowdhury et al., 2007***). The profiles of decapping factors resemble those of Xrn1, suggesting formation of a complex with this 5′→3′ exonuclease or a fast mRNA decay by Xrn1 from 5′→3′ and slowing down towards the 3′ end. Complex formation of Xrn1 and the decapping factors is consistent with a function of Xrn1 in the buffering of mRNA levels in cells (***Sun et al., 2013***), which may be explained if Xrn1 is a regulatory component of the decapping complex.

The preferred localization of 5′ decapping and 3′ deadenylation factors near the mRNA 3′ and 5′ end, respectively, seems counterintuitive, but may be explained by formation of an mRNA closed-loop structure by messenger ribonucleoproteins (mRNPs), where 5′ and 3′ ends are in proximity (***Gallie, 1991***). It is possible that mature mRNPs carry decapping factors near their 3′ end, and that upon polyA tail shorting the decapping complex is activated, leading to decapping and rapid RNA degradation from the 5′ end. In this model, decapping would open the RNA closed-loop structure, providing access for exonucleases and allowing for rapid RNA removal. Further, our result that decapping factors are enriched on translationally non-optimal codons agrees with previous findings that suggest a link between translation and RNA degradation from the 5′ end through the decapping enhancer Dhh1 (***Radhakrishnan et al., 2016***). We note however that our approach does not detect binding events within the polyA tail, limiting further insights.

Comparison of our data with previous profiles of the nuclear surveillance factors Nrd1 and Nab3 reveals many factors that show a similar binding to aberrant non-coding nuclear RNAs, in particular antisense RNA upstream of known promoters, suggesting that these factors are part of the nuclear surveillance machinery. These results are consistent with published data (***Schmid and Jensen, 2018***) and with the following model for nuclear surveillance. First, the Nrd1/Nab3 complex recognizes aberrant, antisense RNAs (***Schulz et al., 2013***). These RNAs then get adenylated by the TRAMP4 complex, consistent with an interaction of the Nrd1/Nab3 complex and the TRAMP subunit Trf4 (***Tudek et al., 2014***). The short polyA tail targets the RNA for degradation by the nuclear exosome. We note that the degradation of introns and ncRNAs upstream of mRNAs on the same strand, which were annotated as NUTs and CUTs, is likely to use the same degradation mechanism because these are bound by the same factors (***Figure 2***, ***Figure 3—figure supplement 2***). Thus, our results are consistent with the idea that degradation of short-lived ncRNAs in the nucleus involves Nrd1/Nab3, TRAMP4, and the exosome.

The exosome is involved in 3′→5′ cytoplasmic mRNA degradation, and in the processing and degradation of long-lived transcripts such as tRNAs, rRNAs and sn(o)RNAs. The differences in

exosome co-factor binding to different RNA classes (***Figure 2***) support the hypothesis that these factors confer specificity for processing and degradation of various RNA species (***Delan-Forino et al., 2017***). The exosome co-factor Ski2 shows cross-linking towards the 3′ end of mRNAs (***Figure 3***). This indicates that the subunit is important for initial RNA degradation by using its helicase activity to dissolve secondary structures and allowing the exosome to start degradation of the transcript (***Schneider and Tollervey, 2013***). Like the exosome complex, the exosome co-factor TRAMP5 binds mRNAs towards the 5′ end of transcripts, and thus some mRNAs may be targeted by TRAMP5 for exosomal degradation. The catalytic exosome subunits Rrp44 and Rrp6 and the exosome core cross-link near the mRNA 5′ end, probably because the exosome moves rapidly from 3′→5′ and then slows down, causing extensive cross-linking. In summary, our resource of transcriptome-binding profiles for 30 RNA degradation factors reveals several hypotheses for the function of these factors that can be tested case by case in the future.

Materials and methods

Key resources table

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Strain, strain background (*S. cerevisiae*, BY4741)	Ccr4_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000000019	
Strain, strain background (*S. cerevisiae*, BY4741)	Pop2_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000005335	
Strain, strain background (*S. cerevisiae*, BY4741)	Not1_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000000689	
Strain, strain background (*S. cerevisiae*, BY4741)	Caf40_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000005232	
Strain, strain background (*S. cerevisiae*, BY4741)	Pan2_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000003062	
Strain, strain background (*S. cerevisiae*, BY4741)	Pan3_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000001508	
Strain, strain background (*S. cerevisiae*, BY4741)	Dcp1_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000005509	
Strain, strain background (*S. cerevisiae*, BY4741)	Dcp2_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000005062	
Strain, strain background (*S. cerevisiae*, BY4741)	Edc2_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000000837	
Strain, strain background (*S. cerevisiae*, BY4741)	Edc3_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000000741	
Strain, strain background (*S. cerevisiae*, BY4741)	Dhh1_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000002319	
Strain, strain background (*S. cerevisiae*, BY4741)	Xrn1_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000003141	
Strain, strain background (*S. cerevisiae*, BY4741)	Rrp6_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000005527	

Continued on next page

Continued

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Strain, strain background (*S. cerevisiae*, BY4741)	Csl4_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000005176	
Strain, strain background (*S. cerevisiae*, BY4741)	Rrp40-TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000005502	
Strain, strain background (*S. cerevisiae*, BY4741)	Rrp4_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000001111	
Strain, strain background (*S. cerevisiae*, BY4741)	Rrp44_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000005381	
Strain, strain background (*S. cerevisiae*, BY4741)	Air1_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000001341	
Strain, strain background (*S. cerevisiae*, BY4741)	Trf5_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000005243	
Strain, strain background (*S. cerevisiae*, BY4741)	Mtr4_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000003586	
Strain, strain background (*S. cerevisiae*, BY4741)	Air2_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000002334	
Strain, strain background (*S. cerevisiae*, BY4741)	Trf4_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000005475	
Strain, strain background (*S. cerevisiae*, BY4741)	Ski2_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000004390	
Strain, strain background (*S. cerevisiae*, BY4741)	Ski3_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000006393	
Strain, strain background (*S. cerevisiae*, BY4741)	Ski7_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000005602	
Strain, strain background (*S. cerevisiae*, BY4741)	Ski8_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000003181	
Strain, strain background (*S. cerevisiae*, BY4741)	Upf1_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000004685	
Strain, strain background (*S. cerevisiae*, BY4741)	Upf2_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000001119	
Strain, strain background (*S. cerevisiae*, BY4741)	Upf3_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000003304	
Strain, strain background (*S. cerevisiae*, BY4741)	Nmd4_TAP	C-terminally tagged gene (Open Biosystems, Germany).	SGD: S000004355	
Antibody	IgG	Sigma-Aldrich	Cat#: I5006, RRID:AB_1163659	IP: 0.1 mg per IP

Continued on next page

Continued

Reagent type (species) or resource	Designation	Source or reference	Identifiers	Additional information
Antibody	PAP anti-TAP	Sigma Aldrich	Cat#: P1291, RRID:AB_1079562	WB (1:2000)
Commercial assay or kit	Dynabeads Protein G	Invitrogen	Cat#: 10003D	330 µl per IP
Commercial assay or kit	RNase T1	Thermo Fisher Scientific	Cat#: EN0541	
Commercial assay or kit	Antarctic P hosphatase	NEB	Cat#: M0289S	
Commercial assay or kit	RNase OUT	Invitrogen	Cat#: 10777019	
Commercial assay or kit	T4 Polynucleotide Kinase	Invitrogen	Cat#: EK0032	
Commercial assay or kit	T4 RNA ligase 2, truncated KQ	NEB	Cat#: M0373S	
Commercial assay or kit	T4 RNA ligase 1	NEB	Cat#: M0437M	
Commercial assay or kit	Proteinase K	NEB	Cat#: P8107S	
Commercial assay or kit	SuperScript III RT	Thermo Fisher Scientific	Cat#: 18080093	
Commercial assay or kit	Phusion High-Fidelity PCR Master Mix	Thermo Fisher Scientific	Cat#: F531S	
Chemical compound, drug	4-thiouracil	Carbosynth	Cat#: 591-28-6	1 mM final conc.
Software, algorithm	mockinbird	***Roth and Torkler, 2018***; https://github.com/soedinglab/Degradation_scripts		
Software, algorithm	UMI-tools	***Smith et al., 2017***; DOI: 10.1101/gr.209601.116		
Software, algorithm	Skewer	***Jiang et al., 2014***; DOI: 10.1186/1471-2105-15-182		
Software, algorithm	Bowtie	***Langmead et al., 2009***; DOI: 10.1186/gb-2009-10-3-r25		
Software, algorithm	tSNE	***Van Der Maaten and Hinton, 2008***; DOI: 10.1007/s10479-011-0841-3		

S. cerevisiae strain verification

Saccharomyces cerevisiae BY4741 strains harboring C-terminally tagged genes (Open Biosystems, Germany) were tested for the correctly inserted tag by Western Blotting using the Peroxidase Anti-Peroxidase (PAP; Sigma) antibody and Pierce ECL Western Blotting Substrate (Thermo Fisher Scientific, USA) (data not shown).

PAR-CLIP experiments of *S. cerevisiae* proteins

PAR-CLIP was performed as described (***Baejen et al., 2014***; ***Battaglia et al., 2017***). Briefly, TAP-tagged protein expressing yeast cells were grown in minimal medium (CSM mixture, Formedium, UK) supplemented with 89 µM uracil, 50–100 µM 4-thiouracil (4tU) and 2% glucose at 30˚C to OD_{600} = 0.5. Cells were labeled in 1 mM 4tU final concentration for 4 hr. After labeling, cells were harvested, resuspended in ice-cold PBS and UV irradiated with 10–12 J/cm^2 at a wavelength of 365 nm on ice and continuous shaking. Lysis was performed in lysis buffer (50 mM Tris-HCl pH 7.5, 100 mM NaCl, 0.5% sodium deoxycholate, 0.1% SDS, 0.5% NP-40) by bead beating (FastPrep−24 Instrument, MP Biomedicals, LLC., France) using silica-zirconium beads (Roth, Germany). The cleared lysate was used for immunoprecipitation with rabbit IgG-conjugated Protein G magnetic beads

(Invitrogen, Germany) on a rotating wheel for 4 hr or overnight at 4˚C. Beads were washed in wash buffer (50 mM Tris-HCl pH 7.5, 1 M NaCl, 0.5% sodium deoxycholate, 0.1% SDS, 0.5% NP-40). IP efficiency was controlled with part of the sample by Western Blot as shown in ***Figure 1—figure supplement 2***. Partial digest of the cross-linked RNA was performed with 50 U RNase T1 per mL for 15–25 min at 25˚C. The dephosphorylation reaction was performed in antarctic phosphatase reaction buffer (NEB, Germany) supplemented with 1 U/µL of antarctic phosphatase and 1 U/µL of RNase OUT (Invitrogen) at 37˚C for 30 min. For rephosphyorylation, beads were incubated in T4 PNK reaction buffer A (Invitrogen) with a final concentration of 1 U/µL T4 PNK, 1 U/µL RNase OUT and 1 mM ATP for 1 hr at 37˚C. 3´ adapter ligation was performed in T4 RNA ligase buffer (NEB) with 10 U/µL T4 RNA ligase 2 (KQ) (NEB), 10 µM 3′ adapter (5′ 5rApp-TGGAATTCTCGGGTGCCAAGG-3ddC 3′ (IDT), 1 U/µL RNase OUT, and 15% (w/v) PEG 8000 overnight at 16˚C. 5´ adapter was ligated to the RNA using T4 RNA ligase buffer (NEB) with 6 U/µL T4 RNA ligase 1 (NEB), 10 µM 5′ adapter (5′ 5I nvddT-GUUCAGAGUUCUACAGUCCGACGAUCNNNNN 3′, IDT), 1 mM ATP, 1 U/µL RNase OUT, 5% (v/v) DMSO, and 10% (w/v) PEG 8000 for 4 hr at 25˚C and 1 hr at 37˚C. Beads were boiled in proteinase K buffer (50 mM Tris-HCl pH 7.5, 6.25 mM EDTA, 75 mM NaCl, 1% SDS) at 95˚C for 5 min. Proteinase K digest was performed with 1.5 mg/mL proteinase K (NEB) for 2 hr at 55˚C. RNA was recovered by acidic phenol/chloroform extraction followed by ethanol precipitation in presence of 0.5 µL GlycoBlue (Invitrogen) and 100 µM RT primer (5′ CCTTGGCACCCGAGAATTCCA 3′, IDT). SuperScript III RTase was used for reverse transcription for 1 hr at 44˚C and 1 hr at 55˚C. NEXTflex barcode primer and universal primer were added to cDNA by PCR amplification with Phusion HF master mix (NEB). After PCR amplification, cDNA was purified and size-selected on a 4% E-Gel EX Agarose Gel (Invitrogen). Quantification on an Agilent 2200 TapeStation instrument (Agilent Technologies, Germany) and 50–75 nt single-end sequencing was performed on Illumina sequencers (HiSeq1500, HiSeq2500 and NextSeq550).

PAR-CLIP data pre-processing

Reads from PAR-CLIP experiments with replicates were merged after making sure that all samples showed high Spearman correlation values comparing binding occupancies of replicates on different genes (***Figure 2—figure supplement 2***). Mapping and statistical evaluation of PAR-CLIP experiments was performed using our in-house software mockinbird (***Roth and Torkler, 2018***). In summary, the UMI is removed from the 5´ end with UMI-tools (***Smith et al., 2017***), and the 3´ adapter is trimmed with Skewer (***Jiang et al., 2014***). Reads with traces of the 5´ adapter are discarded. The preprocessed reads are then mapped to the *S. cerevisiae* genome (sacCer3, version 64.2.1). After mapping PCR duplicates are removed with UMI-tools.

We used two alternative approaches for mapping reads using Bowtie (***Langmead et al., 2009***): For all analyses except the 'transcript class enrichment analysis' in ***Figure 2***, reads are uniquely mapped with up to one mismatch. We discard alignments shorter than 20 nt. This stringent mapping ensures that our high confidence PAR-CLIP cross-link sites are originating from correctly mapped reads on the reference genome. For ***Figure 2***, unique mapping would cause the loss of most reads that fall into rRNAs and tRNAs because of duplicated rRNA genes and tRNA isodecoders. For ***Figure 2***, we therefore allowed Bowtie multi-mapping in two regions with –best, –starra options and discarded reads shorter than 30 nt.

T→C transitions directly at the edge of the reads or with a Phred quality score lower than 20 are not considered as signature of protein binding as they suffer from higher technical noise. To obtain high confidence cross-link sites, we set a stringent cutoff of 0.005 for the p-value of cross-link sites and require a minimum coverage of 2 per site. Moreover, if we see the same transition in at least 75% of reads in the input library control (SRA: SRX532381) (***Baejen et al., 2014***), we annotate it as a single nucleotide polymorphism of our lab strain with respect to the genomic reference and remove such sites from our analysis. Finally, the occupancy of a factor on a verified cross-link site is defined as the number of transitions obtained from our PAR-CLIP experiments divided by the concentration of RNAs covering the cross-link site according to the input library control. This control coverage is measured under comparable conditions to PAR-CLIP experiments (***Baejen et al., 2014***). Occupancy values are capped at the 95th percentile. Subsequent analyses were performed using in-house python scripts. Mockinbird configuration files as well as the analysis scripts can be found at https://github.com/soedinglab/Degradation_scripts (copy archived at https://github.com/elifesciences-publications/Degradation_scripts).

Transcript class enrichment

We analyzed the distribution of reads from high confidence cross-link sites over the genome (***Figure 2A***). We presented the sum of reads from 5′ and 3′ UTRs, coding sequences, and introns as the value for mRNAs. Reads that fall within genomic regions not annotated as categories analyzed here are shown with gray. These annotated transcript classes have comparable U-content, making the comparison between fractions of cross-link sites in each category possible (***Figure 2—figure supplement 2***).

For each factor studied here, we defined enrichment scores that represent their preferences for binding to various transcript classes *c*, in comparison to all other factors. We use annotations for rRNA, tRNA, snoRNA, snRNA, coding sequences (CDS), from *S. cerevisiae* genome sacCer3, version 64.2.1. Untranslated regions around coding boundaries (5′ and 3′ UTRs) were annotated based on TIF-seq experiment (***Pelechano et al., 2013***). We selected the most strongly expressed isoform for each gene. We then assigned boundaries to 3′ and 5′ UTRs based on annotated CDS of the same gene. We furthermore used annotations for stable, unannotated transcripts (SUTs), cryptic unstable transcripts (CUTs), and Nrd1-unterminated transcripts (NUTs) (***Neil et al., 2009***; ***Pelechano et al., 2013***; ***Schulz et al., 2013***). We removed overlapping annotations with the following priority list: rRNA, tRNA, snRNA, snoRNA, intron, CDS, UTR, SUT, CUT, NUT. For each factor, we counted the number of high confidence reads falling in each transcript class. We then used the $\log_2$-transformed matrix and normalized it in the following way for both rows and columns to get log enrichment values that sum to zero in both rows and columns. The row- and sum-normalized enrichment score is defined as follows, where $X_{f,c}$ is the number of high confidence reads for factor *f* that fall into transcript class *c*, and $X'_{f,c} = \log_2 X_{f,c}$ (***Figure 2B***):

$$\tilde{X'_{f,c}} = X'_{f,c} - \frac{X'_{f,\circ} X'_{\circ,c}}{X'_{\circ,\circ}}$$

We defined the row and sum averages of $X_{f,c}$,

$$X'_{f,\circ} = 1C \sum_{c=1}^{C} X'_{f,c},$$

$$X'_{\circ,c} = 1F \sum_{f=1}^{F} X'_{f,c},$$

$$X'_{\circ,\circ} = 1FC \sum_{f=1}^{F} \sum_{c=1}^{C} X'_{f,c},$$

F is the number of factors and C is the number of transcript classes (***Figure 2B***). The normalization can be interpreted as subtracting from the log enrichment matrix X′ the first singular component of its singular-value decomposition.

Metagene analysis

We used the most abundant TIF-annotated isoform for mRNAs (***Pelechano et al., 2013***) as a reference. Transcripts longer than 1500 bases are chosen and aligned at their TSS or pA sites. The average occupancy per nucleotide is then calculated based on high confidence cross-link sites of each PAR-CLIP experiment. The profiles are smoothed by a moving average in a 41 nt window and the 95% confidence interval is estimated by 1500 bootstrap sampling iterations over the transcripts. To further denoise the profiles, the cross-link sites falling in snRNAs, rRNAs, and tRNAs are removed. Furthermore, to avoid ambiguous results, we made sure that the profile comes solely from the central gene. To do so, we performed the metagene analysis around the TSS on the sense strand on TIF-annotated mRNAs that have no other mRNA up to 700 bp upstream of their TSS (3193 transcripts in total). Analogously, for sense-strand pA site profiles we used mRNAs that have no nearby genes downstream of their pA site up to 700 bases on the same strand (3193 transcripts in total). For the antisense strand profiles, we applied the same criteria on the opposite strand which left us

with 3076 and 3193 transcripts filtered around TSS and pA site respectively. This ensures that the observed antisense binding does not originate from neighboring or overlapping transcripts on the antisense strand. In both cases we looked at the average occupancy in a window of [±700 nt] around TSS and around pA sites. Occupancies were normalized to the maximum value, which is the background binding level for antisense profiles with no significant cross-linking to the antisense strand (***Figure 3***, ***Figure 4***, ***Figure 3—figure supplement 3***, and ***Figure 4—figure supplement 2***). The same procedure was followed to plot metagene occupancies centered around protein-coding regions and snoRNAs from *S. cerevisiae* genome sacCer3, version 64.2.1 (***Figure 3—figure supplement 1*** and ***Figure 2—figure supplement 1***). Similarly, CRAC coverage profiles of Xrn1, Mtr4, Trf4, and Ski2 (pre-processed as described in ***Tuck and Tollervey, 2013***) were aligned to TIF-annotated transcripts in the same approach as described here (***Figure 1—figure supplement 1***).

Co-occupancy

Co-occupancy measures the tendency of two factors to bind to the same transcripts. Occupancy of a factor on a transcript is defined as the sum of occupancies for all high confidence cross-link sites falling within this transcript. Co-occupancy of two factors is defined as the Pearson correlation over all transcripts between the occupancies of these factors (***Figure 5A***). We used these correlation values between all pairs of RNA processing factors to assign distances to each pair and used tSNE (***Van Der Maaten and Hinton, 2008***) to visualize the two-dimensional nonlinear embedding of co-occupancies for all RNA-binding proteins in our dataset (***Figure 5C***).

Co-localization

Co-localization measures how likely two factors are to bind near each other in the transcriptome. More precisely, we first calculate the occupancy of a factor $f \in \{1,\ldots,F\}$ around the cross-link sites of another factor f' ([−40 nt,+40 nt] excluding the centered T). We then normalize according to the total occupancy values,

$$z_{ff'} = \sum_{i=1}^{n_{f'}} \left(\sum_{j=-40}^{-1} Occ_{ff',i,j} + \sum_{j=1}^{40} Occ_{ff',i,j} \right)$$

$$\mathrm{co-localization}(f,f') = \frac{z_{ff'}}{\sum_f z_{ff'} \sum_{f'} z_{ff'}}$$

Where, n_f is the number of cross-link sites for factor f, and $Occ_{ff',i,j}$ is the occupancy of f at position j around the i^{th} cross-link site from factor f' ($Occ_{ff',i,j} = 0$ if no verified cross-link sites exist). To improve signal-to-noise, we compute from the resulting matrix of co-localizations between all RNA-processing factors $C_{f,f'}$, the matrix of Pearson correlations between the rows of $C_{f,f'}$, (***Figure 5B***).

Codon-enrichment analysis

To search for possible links between translation efficiency and RNA degradation, we checked if some degradation factors preferentially bind to translationally efficient/non-efficient transcripts. To do so we adapted the proposed normalized translation efficiency scale (***Pechmann and Frydman, 2013***). The authors generate a normalized optimality score for codons that incorporates the competition between supply and demand of tRNAs. The coding region for each transcript was extracted according to ORFs annotated by SGD. The codon optimality score was averaged over the whole reading frame (***Figure 6A***, more detailed explanation in the next section).

We then checked whether mRNAs that bind to each factor are enriched or depleted in some codons compared to all mRNAs. To achieve this, we defined the following score for codon enrichment that represents deviations from average frequencies in all mRNAs,

$$codon\,enrichment = \frac{\sum_{t=1}^{T} \left(\frac{occ(t)}{\sum_{t'=1}^{T} occ(t')} \times F_{c,t} \right)}{\frac{1}{T}\sum_{t=1}^{T} F_{c,t}}$$

Here T is the number of mRNA transcripts, $F_{c,t}$ is the fraction of the codon c in transcript t, and $occ(t)$ is the *total* occupancy of the factor on transcript t. 90% confidence intervals were

generated by bootstrapping: we sampled *with replacement* 1000 times the same number of mRNAs from the total set as in total, and for each set we recalculated the codon enrichment score. We colored the bars based on the previously ranked optimality of codons (**Pechmann and Frydman, 2013**) (**Figure 6A**, **Figure 6—figure supplements 1–7**).

Relating occupancies to various transcript features

We analyzed the correlation of the occupancy of all factors with transcript length, codon enrichment of the transcript, expression level, transcript stability, and polyA tail length. For expression, we used an RNA-seq experiment of wild-type yeast (SRA: SRX532381) (**Baejen et al., 2017**) and mapped the reads to mRNAs. We present the average number of reads per base as an estimate for gene expression. For half-life calculations, we used published yeast 4tU-seq (GEO: GSM2199309) and RNA-seq experiments (SRA: SRX532381) (**Baejen et al., 2017**). Transcript half-life is estimated with an optimized method that will be published elsewhere (Hofmann et al., unpublished).

Since there are only few transcripts with very low or very high half-life, codon optimality, and expression (**Figure 6—figure supplement 8**), we performed the analysis on a subset of mRNAs where the transcript property lies between the 5% and 95% quantiles. We then compared the *total* occupancy of degradation factors on each mRNA relative to such transcript features (**Figures 6A** and **7B**, and **Figure 6—figure supplements 1–7**). We show 95% confidence intervals generated by bootstrapping mRNAs in gray shade.

We checked whether such correlations originate from the feature of interest or merely shows up due to correlations between this feature and others (**Figure 6—figure supplement 8**). We used a multivariate linear regression to model total occupancy as a linear function of these four features:

$$occupancy'(t) \sim length + optimality + expression + half\ life$$

In cases where the correlation is a direct effect from our feature of interest, we expect to lose significantly on our prediction when this variable is taken out of the equation. Therefore, we use p-values representing the importance of each feature in this linear regression as a score representing the significance of its contribution in explaining the final occupancies. Occupancy correlated strongly with transcript length, which dominated as explanatory variable in this regression, trivially because most factors bind along the entire transcript. To eliminate this trivial dependency, we used occupancy per nucleotide, denoted $occupancy'$, as the target variable in our regression (**Figure 6C**).

Motif enrichment analysis

To find sequence preferences for binding events of degradation factors, we counted 4-mers in a window of [±5 nt] intervals around high confidence cross-link sites of PAR-CLIP experiments. Based on this count table, the enrichment score for each 4-mer was calculated using the following formula,

$$enrichment\,(4-\mathrm{mer}, i) = \frac{n_{4-\mathrm{mer},i} + 1}{N \times \Pi_{j=1}^{4} P_{4-\mathrm{mer[j]}}}$$

Here N is the number of cross-link sites below the cut-off p-value (we used a maximum of 5000 cross link sites), $n_{4mer,\,i}$ is the number of observed *4-mers* at position i in the set of binding sequences aligned at their cross-link site i=0, 4-*mer*[j] is the base at the j'th position of the 4-mer, and P_b is the probability of observing base b. We used the probabilities: $P_A = P_T = 0.31$ and $P_C = P_G = 0.19$ based on frequencies in yeast genome and corrected for the T bias at the cross-link site (**Figure 4—figure supplement 1**).

Acknowledgements

We would like to thank Helmut Blum (LAFUGA, LMU Munich), Stefan Krebs (LAFUGA, LMU Munich), Kerstin Maier and Petra Rus (Cramer laboratory) for sequencing. We thank Gabriel Villamil and Bjoern Schwalb (Cramer laboratory) for sharing RNA half-life calculations prior to publication. PC was funded by the Advanced Grant TRANSREGULON of the European Research Council and by the Volkswagen Foundation. This work was supported by the DFG SPP1935 grant CR 117/6–1.

Additional information

Funding

Funder	Grant reference number	Author
European Research Council	Advanced Grant Transregulon	Patrick Cramer
Volkswagen Foundation		Patrick Cramer
Deutsche Forschungsgemeinschaft	SPP1935 grant CR 117/6-1	Johannes Soeding Patrick Cramer
Max-Planck-Gesellschaft	Open-Access funding	Patrick Cramer

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

Salma Sohrabi-Jahromi, Formal analysis, Validation, Investigation, Visualization, Methodology, Writing—original draft, Writing—review and editing; Katharina B Hofmann, Data curation, Validation, Investigation, Methodology, Writing—original draft, Writing—review and editing; Andrea Boltendahl, Saskia Gressel, Carlo Baejen, Data curation, Methodology; Christian Roth, Formal analysis, Visualization, Methodology; Johannes Soeding, Conceptualization, Supervision, Funding acquisition, Methodology, Writing—original draft, Project administration, Writing—review and editing; Patrick Cramer, Conceptualization, Supervision, Funding acquisition, Writing—original draft, Project administration, Writing—review and editing

Author ORCIDs

Salma Sohrabi-Jahromi https://orcid.org/0000-0002-8417-8230
Katharina B Hofmann https://orcid.org/0000-0002-0683-6277
Saskia Gressel http://orcid.org/0000-0003-0261-675X
Patrick Cramer https://orcid.org/0000-0001-5454-7755

Decision letter and Author response

Decision letter https://doi.org/10.7554/eLife.47040.044
Author response https://doi.org/10.7554/eLife.47040.045

Additional files

Supplementary files

- Supplementary file 1. Overview of RNA processing factors and their respective PAR-CLIP experiments used in this study.

DOI: https://doi.org/10.7554/eLife.47040.031

- Transparent reporting form

DOI: https://doi.org/10.7554/eLife.47040.032

Data availability

Sequencing data have been deposited in GEO under accession codes GSE 128312.

The following dataset was generated:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Sohrabi-Johromi S, Hofmann KB, Boltendahl A, Roth C, Gressel S, Baejen C, Soeding J, Cramer P	2019	Transcriptome maps of general eukaryotic RNA degradation factors	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE128312	NCBI Gene Expression Omnibus, GSE8128312

The following previously published datasets were used:

Author(s)	Year	Dataset title	Dataset URL	Database and Identifier
Schulz D, Schwalb B, Kiesel A, Baejen C, Torkler P, Gagneur J, Soeding J, Cramer P	2013	Nuclear depletion of the essential transcription termination factor Nrd1 in Saccharomyces cerevisiae was studied using a combination of RNA-Seq, ChIP-Seq of Pol II and PAR-CLIP of Nrd1	https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-1766/	Array Express, E-MTAB-1766
Baejen C, Torkler P, Gressel S, Essig K, Söding J, Cramer P	2014	Transcriptome maps of mRNP biogenesis factors define pre-mRNA recognition	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE59676	NCBI Gene Expression Omnibus, GSE59676
Baejen C, Andreani J, Torkler P, Battaglia S, Schwalb B, Lidschreiber M, Maier KC, Boltendahl A, Rus P, Esslinger S, Soeding J, Cramer P	2017	Genome-wide analysis of RNA polymerase II termination at protein-coding genes.	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE79222	NCBI Gene Expression Omnibus, GSE79222
Battaglia S, Lidschreiber M, Baejen C, Torkler P, Vos S, Cramer P	2017	RNA-dependent chromatin association of transcription elongation factors and Pol II CTD kinases	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81822	NCBI Gene Expression Omnibus, GSE81822

References

Allmang C, Petfalski E, Podtelejnikov A, Mann M, Tollervey D, Mitchell P. 1999. The yeast exosome and human PM-Scl are related complexes of 3′ –> 5′ exonucleases. *Genes & Development* **13**:2148–2158. DOI: https://doi.org/10.1101/gad.13.16.2148, PMID: 10465791

Anderson JS, Parker RP. 1998. The 3′ to 5′ degradation of yeast mRNAs is a general mechanism for mRNA turnover that requires the SKI2 DEVH box protein and 3′ to 5′ exonucleases of the exosome complex. *The EMBO Journal* **17**:1497–1506. DOI: https://doi.org/10.1093/emboj/17.5.1497, PMID: 9482746

Araki Y, Takahashi S, Kobayashi T, Kajiho H, Hoshino S, Katada T. 2001. Ski7p G protein interacts with the exosome and the Ski complex for 3′-to-5′ mRNA decay in yeast. *The EMBO Journal* **20**:4684–4693. DOI: https://doi.org/10.1093/emboj/20.17.4684, PMID: 11532933

Baejen C, Torkler P, Gressel S, Essig K, Söding J, Cramer P. 2014. Transcriptome maps of mRNP biogenesis factors define pre-mRNA recognition. *Molecular Cell* **55**:745–757. DOI: https://doi.org/10.1016/j.molcel.2014.08.005, PMID: 25192364

Baejen C, Andreani J, Torkler P, Battaglia S, Schwalb B, Lidschreiber M, Maier KC, Boltendahl A, Rus P, Esslinger S, Söding J, Cramer P. 2017. Genome-wide analysis of RNA polymerase II termination at Protein-Coding genes. *Molecular Cell* **66**:38–49. DOI: https://doi.org/10.1016/j.molcel.2017.02.009, PMID: 28318822

Baker KE, Parker R. 2004. Nonsense-mediated mRNA decay: terminating erroneous gene expression. *Current Opinion in Cell Biology* **16**:293–299. DOI: https://doi.org/10.1016/j.ceb.2004.03.003, PMID: 15145354

Battaglia S, Lidschreiber M, Baejen C, Torkler P, Vos SM, Cramer P. 2017. RNA-dependent chromatin association of transcription elongation factors and pol II CTD kinases. *eLife* **6**:e25637. DOI: https://doi.org/10.7554/eLife.25637, PMID: 28537551

Beilharz TH, Preiss T. 2007. Widespread use of poly(A) tail length control to accentuate expression of the yeast transcriptome. *RNA* **13**:982–997. DOI: https://doi.org/10.1261/rna.569407, PMID: 17586758

Boeck R, Tarun S, Rieger M, Deardorff JA, Müller-Auer S, Sachs AB. 1996. The yeast Pan2 protein is required for poly(A)-binding protein-stimulated poly(A)-nuclease activity. *Journal of Biological Chemistry* **271**:432–438. DOI: https://doi.org/10.1074/jbc.271.1.432, PMID: 8550599

Bonneau F, Basquin J, Ebert J, Lorentzen E, Conti E. 2009. The yeast exosome functions as a macromolecular cage to channel RNA substrates for degradation. *Cell* **139**:547–559. DOI: https://doi.org/10.1016/j.cell.2009.08.042, PMID: 19879841

Brown JT, Bai X, Johnson AW. 2000. The yeast antiviral proteins Ski2p, Ski3p, and Ski8p exist as a complex in vivo. *RNA* **6**:449–457. DOI: https://doi.org/10.1017/S1355838200991787, PMID: 10744028

Brown CE, Sachs AB. 1998. Poly(A) tail length control in Saccharomyces cerevisiae occurs by message-specific deadenylation. *Molecular and Cellular Biology* **18**:6548–6559. DOI: https://doi.org/10.1128/MCB.18.11.6548, PMID: 9774670

Cao D, Parker R. 2003. Computational modeling and experimental analysis of nonsense-mediated decay in yeast. *Cell* **113**:533–545. DOI: https://doi.org/10.1016/S0092-8674(03)00353-2, PMID: 12757713

Caponigro G, Parker R. 1995. Multiple functions for the poly(A)-binding protein in mRNA decapping and deadenylation in yeast. *Genes & Development* **9**:2421–2432. DOI: https://doi.org/10.1101/gad.9.19.2421, PMID: 7557393

Celik A, Baker R, He F, Jacobson A. 2017. High-resolution profiling of NMD targets in yeast reveals translational fidelity as a basis for substrate selection. *RNA* **23**:735–748. DOI: https://doi.org/10.1261/rna.060541.116, PMID: 28209632

Chakrabarti S, Jayachandran U, Bonneau F, Fiorini F, Basquin C, Domcke S, Le Hir H, Conti E. 2011. Molecular mechanisms for the RNA-dependent ATPase activity of Upf1 and its regulation by Upf2. *Molecular Cell* **41**:693–703. DOI: https://doi.org/10.1016/j.molcel.2011.02.010, PMID: 21419344

Chowdhury A, Mukhopadhyay J, Tharun S. 2007. The decapping activator Lsm1p-7p-Pat1p complex has the intrinsic ability to distinguish between oligoadenylated and polyadenylated RNAs. *RNA* **13**:998–1016. DOI: https://doi.org/10.1261/rna.502507, PMID: 17513695

Delan-Forino C, Schneider C, Tollervey D. 2017. Transcriptome-wide analysis of alternative routes for RNA substrates into the exosome complex. *PLOS Genetics* **13**:e1006699. DOI: https://doi.org/10.1371/journal.pgen.1006699, PMID: 28355211

Doma MK, Parker R. 2007. RNA quality control in eukaryotes. *Cell* **131**:660–668. DOI: https://doi.org/10.1016/j.cell.2007.10.041, PMID: 18022361

Finoux AL, Séraphin B. 2006. In vivo targeting of the yeast Pop2 deadenylase subunit to reporter transcripts induces their rapid degradation and generates new decay intermediates. *Journal of Biological Chemistry* **281**:25940–25947. DOI: https://doi.org/10.1074/jbc.M600132200, PMID: 16793769

Franks TM, Singh G, Lykke-Andersen J. 2010. Upf1 ATPase-dependent mRNP disassembly is required for completion of nonsense- mediated mRNA decay. *Cell* **143**:938–950. DOI: https://doi.org/10.1016/j.cell.2010.11.043, PMID: 21145460

Gallie DR. 1991. The cap and poly(A) tail function synergistically to regulate mRNA translational efficiency. *Genes & Development* **5**:2108–2116. DOI: https://doi.org/10.1101/gad.5.11.2108, PMID: 1682219

Geisberg JV, Moqtaderi Z, Fan X, Ozsolak F, Struhl K. 2014. Global analysis of mRNA isoform half-lives reveals stabilizing and destabilizing elements in yeast. *Cell* **156**:812–824. DOI: https://doi.org/10.1016/j.cell.2013.12.026, PMID: 24529382

Goldstrohm AC, Hook BA, Seay DJ, Wickens M. 2006. PUF proteins bind Pop2p to regulate messenger RNAs. *Nature Structural & Molecular Biology* **13**:533–539. DOI: https://doi.org/10.1038/nsmb1100, PMID: 16715093

Grzechnik P, Kufel J. 2008. Polyadenylation linked to transcription termination directs the processing of snoRNA precursors in yeast. *Molecular Cell* **32**:247–258. DOI: https://doi.org/10.1016/j.molcel.2008.10.003, PMID: 18951092

Heck AM, Wilusz J. 2018. The interplay between the RNA decay and translation machinery in eukaryotes. *Cold Spring Harbor Perspectives in Biology* **10**:a032839. DOI: https://doi.org/10.1101/cshperspect.a032839, PMID: 29311343

Heo DH, Yoo I, Kong J, Lidschreiber M, Mayer A, Choi BY, Hahn Y, Cramer P, Buratowski S, Kim M. 2013. The RNA polymerase II C-terminal domain-interacting domain of yeast Nrd1 contributes to the choice of termination pathway and couples to RNA processing by the nuclear exosome. *Journal of Biological Chemistry* **288**:36676–36690. DOI: https://doi.org/10.1074/jbc.M113.508267, PMID: 24196955

Houseley J, LaCava J, Tollervey D. 2006. RNA-quality control by the exosome. *Nature Reviews Molecular Cell Biology* **7**:529–539. DOI: https://doi.org/10.1038/nrm1964, PMID: 16829983

Houseley J, Tollervey D. 2009. The many pathways of RNA degradation. *Cell* **136**:763–776. DOI: https://doi.org/10.1016/j.cell.2009.01.019, PMID: 19239894

Huch S, Nissan T. 2014. Interrelations between translation and general mRNA degradation in yeast. *Wiley Interdisciplinary Reviews: RNA* **5**:747–763. DOI: https://doi.org/10.1002/wrna.1244, PMID: 24944158

Jiang H, Lei R, Ding SW, Zhu S. 2014. Skewer: a fast and accurate adapter trimmer for next-generation sequencing paired-end reads. *BMC Bioinformatics* **15**:182. DOI: https://doi.org/10.1186/1471-2105-15-182, PMID: 24925680

Jiao X, Xiang S, Oh C, Martin CE, Tong L, Kiledjian M. 2010. Identification of a quality-control mechanism for mRNA 5'-end capping. *Nature* **467**:608–611. DOI: https://doi.org/10.1038/nature09338, PMID: 20802481

Kruk JA, Dutta A, Fu J, Gilmour DS, Reese JC. 2011. The multifunctional Ccr4-Not complex directly promotes transcription elongation. *Genes & Development* **25**:581–593. DOI: https://doi.org/10.1101/gad.2020911, PMID: 21406554

Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology* **10**:R25. DOI: https://doi.org/10.1186/gb-2009-10-3-r25, PMID: 19261174

Lebreton A, Séraphin B. 2008. Exosome-mediated quality control: substrate recruitment and molecular activity. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1779**:558–565. DOI: https://doi.org/10.1016/j.bbagrm.2008.02.003, PMID: 18313413

Losson R, Lacroute F. 1979. Interference of nonsense mutations with eukaryotic messenger RNA stability. *PNAS* **76**:5134–5137. DOI: https://doi.org/10.1073/pnas.76.10.5134, PMID: 388431

Lykke-Andersen S, Brodersen DE, Jensen TH. 2009. Origins and activities of the eukaryotic exosome. *Journal of Cell Science* **122**:1487–1494. DOI: https://doi.org/10.1242/jcs.047399, PMID: 19420235

Marquardt S, Hazelbaker DZ, Buratowski S. 2011. Distinct RNA degradation pathways and 3' extensions of yeast non-coding RNA species. *Transcription* **2**:145–154. DOI: https://doi.org/10.4161/trns.2.3.16298, PMID: 21826286

Miller C, Schwalb B, Maier K, Schulz D, Dümcke S, Zacher B, Mayer A, Sydow J, Marcinowski L, Dölken L, Martin DE, Tresch A, Cramer P. 2011. Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Molecular Systems Biology* **7**:458. DOI: https://doi.org/10.1038/msb.2010.112, PMID: 21206491

Milligan L, Decourty L, Saveanu C, Rappsilber J, Ceulemans H, Jacquier A, Tollervey D. 2008. A yeast exosome cofactor, Mpp6, functions in RNA surveillance and in the degradation of noncoding RNA transcripts. *Molecular and Cellular Biology* **28**:5446–5457. DOI: https://doi.org/10.1128/MCB.00463-08, PMID: 18591258

Milligan L, Huynh-Thu VA, Delan-Forino C, Tuck A, Petfalski E, Lombraña R, Sanguinetti G, Kudla G, Tollervey D. 2016. Strand-specific, high-resolution mapping of modified RNA polymerase II. *Molecular Systems Biology* **12**: 874. DOI: https://doi.org/10.15252/msb.20166869, PMID: 27288397

Mitchell P, Petfalski E, Houalla R, Podtelejnikov A, Mann M, Tollervey D. 2003. Rrp47p is an exosome-associated protein required for the 3' processing of stable RNAs. *Molecular and Cellular Biology* **23**:6982–6992. DOI: https://doi.org/10.1128/MCB.23.19.6982-6992.2003, PMID: 12972615

Mitchell P, Tollervey D. 2003. An NMD pathway in yeast involving accelerated deadenylation and exosome-mediated 3'–>5' degradation. *Molecular Cell* **11**:1405–1413. DOI: https://doi.org/10.1016/S1097-2765(03)00190-4, PMID: 12769863

Morrissey JP, Deardorff JA, Hebron C, Sachs AB. 1999. Decapping of stabilized, polyadenylated mRNA in yeast pab1 mutants. *Yeast* **15**:687–702. DOI: https://doi.org/10.1002/(SICI)1097-0061(19990615)15:8<687::AID-YEA412>3.0.CO;2-L, PMID: 10392446

Muhlrad D, Parker R. 1994. Premature translational termination triggers mRNA decapping. *Nature* **370**:578–581. DOI: https://doi.org/10.1038/370578a0, PMID: 8052314

Neil H, Malabat C, d'Aubenton-Carafa Y, Xu Z, Steinmetz LM, Jacquier A. 2009. Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature* **457**:1038–1042. DOI: https://doi.org/10.1038/nature07747, PMID: 19169244

Ogami K, Chen Y, Manley JL. 2018. RNA surveillance by the nuclear RNA exosome: mechanisms and significance. *Non-Coding RNA* **4**:8. DOI: https://doi.org/10.3390/ncrna4010008, PMID: 29629374

Parker R. 2012. RNA degradation in Saccharomyces cerevisae. *Genetics* **191**:671–702. DOI: https://doi.org/10.1534/genetics.111.137265, PMID: 22785621

Pechmann S, Frydman J. 2013. Evolutionary conservation of codon optimality reveals hidden signatures of cotranslational folding. *Nature Structural & Molecular Biology* **20**:237–243. DOI: https://doi.org/10.1038/nsmb.2466, PMID: 23262490

Pelechano V, Wei W, Steinmetz LM. 2013. Extensive transcriptional heterogeneity revealed by isoform profiling. *Nature* **497**:127–131. DOI: https://doi.org/10.1038/nature12121, PMID: 23615609

Radhakrishnan A, Chen YH, Martin S, Alhusaini N, Green R, Coller J. 2016. The DEAD-Box protein Dhh1p couples mRNA decay and translation by monitoring codon optimality. *Cell* **167**:122–132. DOI: https://doi.org/10.1016/j.cell.2016.08.053, PMID: 27641505

Roth C, Torkler P. 2018. *Soedinglab/mockinbird: Degradation PAR-CLIP*. f331095. GitHub. https://github.com/soedinglab/Degradation_scripts

Schaeffer D, Tsanova B, Barbas A, Reis FP, Dastidar EG, Sanchez-Rotunno M, Arraiano CM, van Hoof A. 2009. The exosome contains domains with specific endoribonuclease, exoribonuclease and cytoplasmic mRNA decay activities. *Nature Structural & Molecular Biology* **16**:56–62. DOI: https://doi.org/10.1038/nsmb.1528, PMID: 19060898

Schmid M, Jensen TH. 2018. Controlling nuclear RNA levels. *Nature Reviews Genetics* **19**:518–529. DOI: https://doi.org/10.1038/s41576-018-0013-2, PMID: 29748575

Schneider C, Kudla G, Wlotzka W, Tuck A, Tollervey D. 2012. Transcriptome-wide analysis of exosome targets. *Molecular Cell* **48**:422–433. DOI: https://doi.org/10.1016/j.molcel.2012.08.013, PMID: 23000172

Schneider C, Tollervey D. 2013. Threading the barrel of the RNA exosome. *Trends in Biochemical Sciences* **38**: 485–493. DOI: https://doi.org/10.1016/j.tibs.2013.06.013, PMID: 23910895

Schulz D, Schwalb B, Kiesel A, Baejen C, Torkler P, Gagneur J, Soeding J, Cramer P. 2013. Transcriptome surveillance by selective termination of noncoding RNA synthesis. *Cell* **155**:1075–1087. DOI: https://doi.org/10.1016/j.cell.2013.10.024, PMID: 24210918

Semotok JL, Cooperstock RL, Pinder BD, Vari HK, Lipshitz HD, Smibert CA. 2005. Smaug recruits the CCR4/POP2/NOT deadenylase complex to trigger maternal transcript localization in the early Drosophila embryo. *Current Biology* **15**:284–294. DOI: https://doi.org/10.1016/j.cub.2005.01.048, PMID: 15723788

Smith JE, Alvarez-Dominguez JR, Kline N, Huynh NJ, Geisler S, Hu W, Coller J, Baker KE. 2014. Translation of small open reading frames within unannotated RNA transcripts in Saccharomyces cerevisiae. *Cell Reports* **7**: 1858–1866. DOI: https://doi.org/10.1016/j.celrep.2014.05.023, PMID: 24931603

Smith T, Heger A, Sudbery I. 2017. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Research* **27**:491–499. DOI: https://doi.org/10.1101/gr.209601.116, PMID: 28100584

Stevens A, Poole TL. 1995. 5'-exonuclease-2 of saccharomyces cerevisiae. Purification and features of ribonuclease activity with comparison to 5'-exonuclease-1. *The Journal of Biological Chemistry* **270**:16063–16069. DOI: https://doi.org/10.1074/jbc.270.27.16063, PMID: 7608167

Sun M, Schwalb B, Pirkl N, Maier KC, Schenk A, Failmezger H, Tresch A, Cramer P. 2013. Global analysis of eukaryotic mRNA degradation reveals Xrn1-dependent buffering of transcript levels. *Molecular Cell* **52**:52–62. DOI: https://doi.org/10.1016/j.molcel.2013.09.010, PMID: 24119399

Synowsky SA, van Wijk M, Raijmakers R, Heck AJ. 2009. Comparative multiplexed mass spectrometric analyses of endogenously expressed yeast nuclear and cytoplasmic exosomes. *Journal of Molecular Biology* **385**:1300–1313. DOI: https://doi.org/10.1016/j.jmb.2008.11.011, PMID: 19046973

Tharun S, Parker R. 2001. Targeting an mRNA for decapping: displacement of translation factors and association of the Lsm1p-7p complex on deadenylated yeast mRNAs. *Molecular Cell* **8**:1075–1083. DOI: https://doi.org/10.3410/f.1002688.28954, PMID: 11741542

Thompson DM, Parker R. 2007. Cytoplasmic decay of intergenic transcripts in Saccharomyces cerevisiae. *Molecular and Cellular Biology* **27**:92–101. DOI: https://doi.org/10.1128/MCB.01023-06, PMID: 17074811

Tuck AC, Tollervey D. 2013. A transcriptome-wide atlas of RNP composition reveals diverse classes of mRNAs and lncRNAs. *Cell* **154**:996–1009. DOI: https://doi.org/10.1016/j.cell.2013.07.047, PMID: 23993093

Tucker M, Valencia-Sanchez MA, Staples RR, Chen J, Denis CL, Parker R. 2001. The transcription factor associated Ccr4 and Caf1 proteins are components of the major cytoplasmic mRNA deadenylase in Saccharomyces cerevisiae. *Cell* **104**:377–386. DOI: https://doi.org/10.1016/S0092-8674(01)00225-2, PMID: 11239395

Tudek A, Porrua O, Kabzinski T, Lidschreiber M, Kubicek K, Fortova A, Lacroute F, Vanacova S, Cramer P, Stefl R, Libri D. 2014. Molecular basis for coordinating transcription termination with noncoding RNA degradation. *Molecular Cell* **55**:467–481. DOI: https://doi.org/10.1016/j.molcel.2014.05.031, PMID: 25066235

Turowski TW, Tollervey D. 2015. Cotranscriptional events in eukaryotic ribosome synthesis. *Wiley Interdisciplinary Reviews: RNA* **6**:129–139. DOI: https://doi.org/10.1002/wrna.1263, PMID: 25176256

Van Der Maaten LJP, Hinton GE. 2008. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research : JMLR* **9**:2579–2605.

van Hoof A, Frischmeyer PA, Dietz HC, Parker R. 2002. Exosome-mediated recognition and degradation of mRNAs lacking a termination codon. *Science* **295**:2262–2264. DOI: https://doi.org/10.1126/science.1067272, PMID: 11910110

Vanácová S, Wolf J, Martin G, Blank D, Dettwiler S, Friedlein A, Langen H, Keith G, Keller W. 2005. A new yeast poly(A) polymerase complex involved in RNA quality control. *PLOS Biology* **3**:e189. DOI: https://doi.org/10.1371/journal.pbio.0030189, PMID: 15828860

Vasiljeva L, Buratowski S. 2006. Nrd1 interacts with the nuclear exosome for 3' processing of RNA polymerase II transcripts. *Molecular Cell* **21**:239–248. DOI: https://doi.org/10.1016/j.molcel.2005.11.028, PMID: 16427013

Vicens Q, Kieft JS, Rissland OS. 2018. Revisiting the Closed-Loop model and the nature of mRNA 5'-3' Communication. *Molecular Cell* **72**:805–812. DOI: https://doi.org/10.1016/j.molcel.2018.10.047, PMID: 30526871

Wang L, Lewis MS, Johnson AW. 2005. Domain interactions within the Ski2/3/8 complex and between the Ski complex and Ski7p. *RNA* **11**:1291–1302. DOI: https://doi.org/10.1261/rna.2060405, PMID: 16043509

Webster MW, Chen YH, Stowell JAW, Alhusaini N, Sweet T, Graveley BR, Coller J, Passmore LA. 2018. mRNA deadenylation is coupled to translation rates by the differential activities of Ccr4-Not nucleases. *Molecular Cell* **70**:1089–1100. DOI: https://doi.org/10.1016/j.molcel.2018.05.033, PMID: 29932902

Wells SE, Hillner PE, Vale RD, Sachs AB. 1998. Circularization of mRNA by eukaryotic translation initiation factors. *Molecular Cell* **2**:135–140. DOI: https://doi.org/10.1016/S1097-2765(00)80122-7, PMID: 9702200

Wolf J, Passmore LA. 2014. mRNA deadenylation by Pan2-Pan3. *Biochemical Society Transactions* **42**:184–187. DOI: https://doi.org/10.1042/BST20130211, PMID: 24450649

Wyers F, Rougemaille M, Badis G, Rousselle JC, Dufour ME, Boulay J, Régnault B, Devaux F, Namane A, Séraphin B, Libri D, Jacquier A. 2005. Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase. *Cell* **121**:725–737. DOI: https://doi.org/10.1016/j.cell.2005.04.030, PMID: 15935759

Zinder JC, Lima CD. 2017. Targeting RNA for processing or destruction by the eukaryotic RNA exosome and its cofactors. *Genes & Development* **31**:88–100. DOI: https://doi.org/10.1101/gad.294769.116, PMID: 28202538

eLIFE elifesciences.org

RESEARCH ARTICLE

Figures and figure supplements

Transcriptome maps of general eukaryotic RNA degradation factors

Salma Sohrabi-Jahromi *et al*

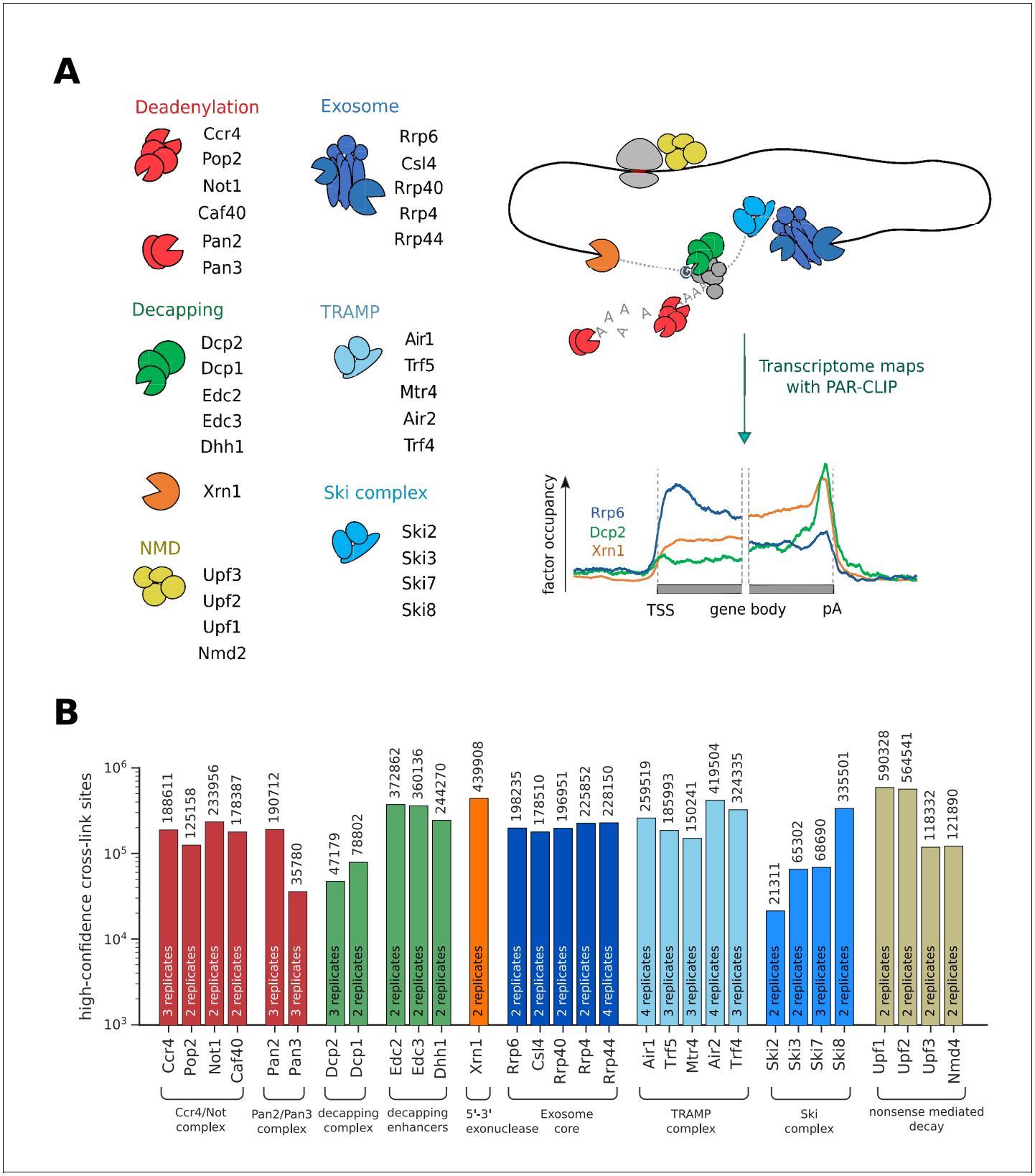


Figure 1. Overview of PAR-CLIP experiments performed in this study. (**A**) Overview of degradation pathways studied. (**B**) Number of high-confidence PAR-CLIP cross-link sites obtained for each factor after merging data from replicates.

DOI: https://doi.org/10.7554/eLife.47040.003

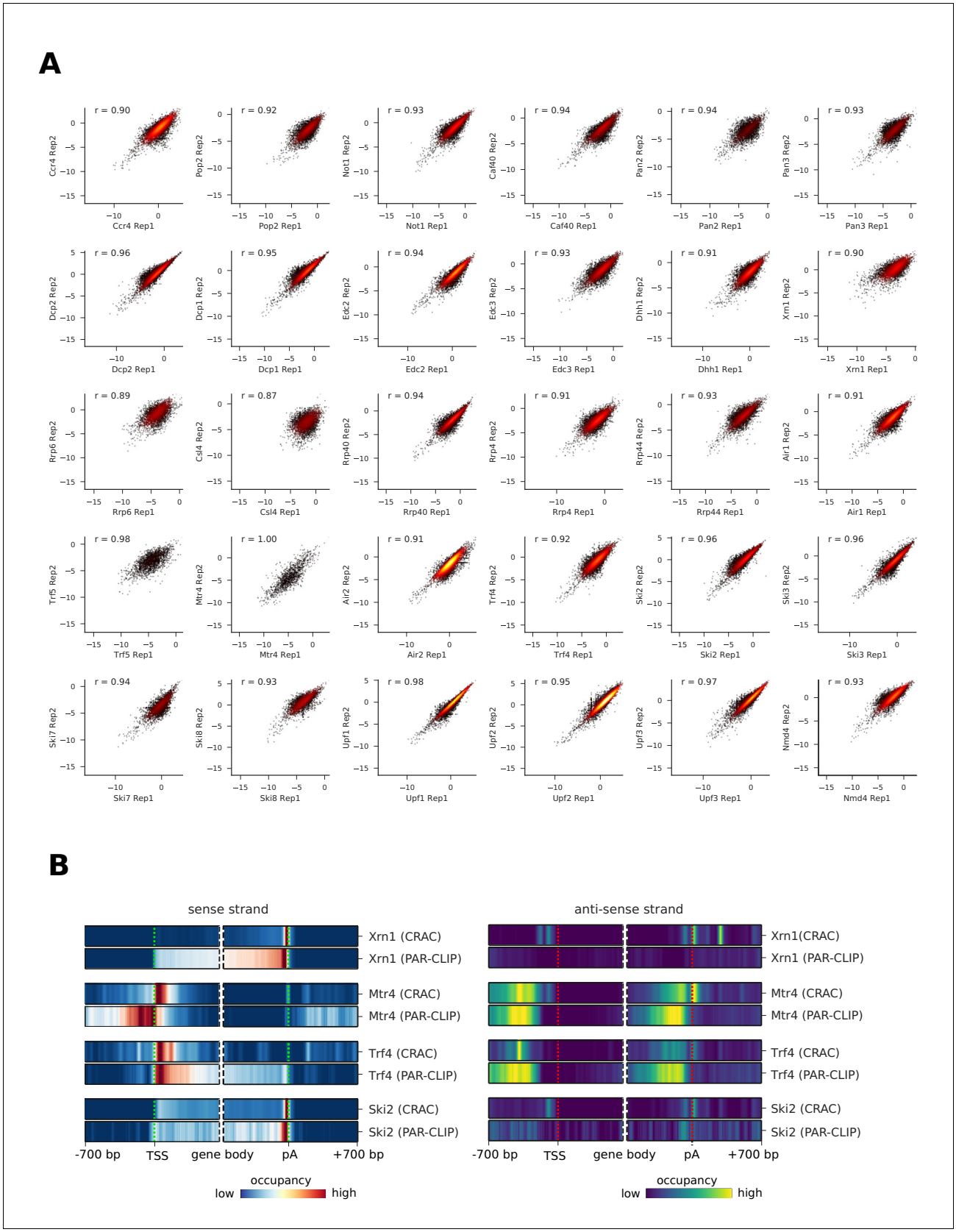


Figure 1—figure supplement 1. Biological replicate PAR-CLIP experiments have high correlation. (**A**) Total transcript occupancy of factors in replicate experiments are plotted (in $\log_2$ space) and Spearman correlation values are shown for each pair. Each dot corresponds to a transcript. The color

Figure 1—figure supplement 1 continued on next page

Figure 1—figure supplement 1 continued

indicates dot density. (**B**) Comparison of coverage profiles obtained from CRAC experiments of Xrn1, Mtr4, Trf4, and Ski2 in *S. cerevisiae* (***Tuck and Tollervey, 2013***) with occupancy profiles from our PAR-CLIP experiments highlights reproducibility of transcriptome profiles across different methods. These profiles show the averaged binding of degradation factors over mRNAs (sense strand: left and anti-sense strand: right) in a window of [±700 nt] around their transcription start site (TSS) and their poly-adenylation (pA) site in a window of [±700 nt]. Regions that have neighboring transcripts on the same strand were removed to avoid contaminating profiles (Materials and methods).

DOI: https://doi.org/10.7554/eLife.47040.004

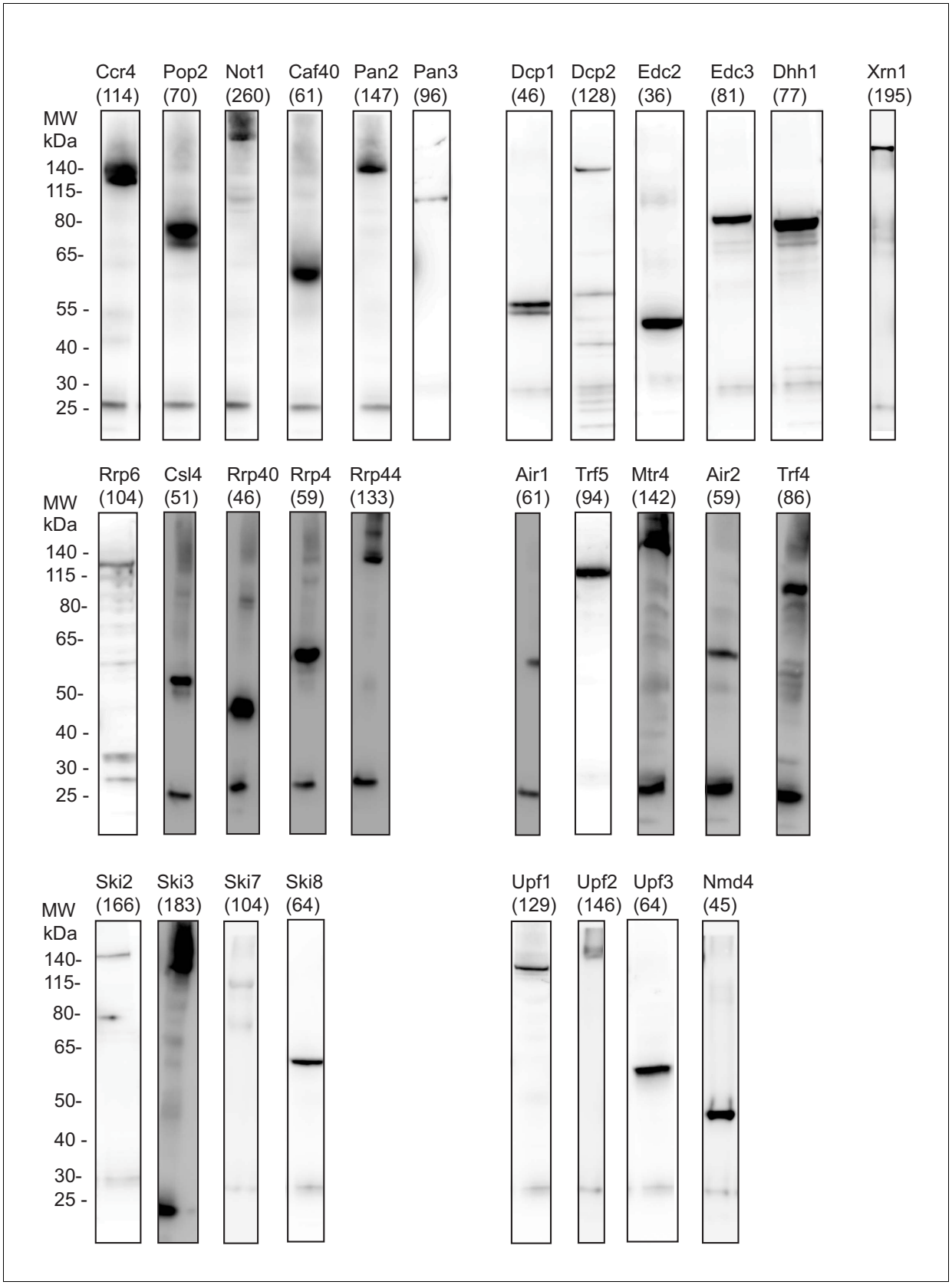


Figure 1—figure supplement 2. Western Blot analysis for all degradation factors analyzed in this study show IP efficiency. IP using the TAP-tag is detected by Western Blot analysis with tag specific antibody to show the IP quality of the different PAR-CLIP experiments representative for one

Figure 1—figure supplement 2 continued on next page

Figure 1—figure supplement 2 continued

replicate per factor. The factors are sorted according to their complexes: deadenylation (Ccr4, Pop2, Not1, Caf40, Pan3, and Pan3), decapping (Dcp1, Dcp2, Edc2, Edc3, and Dhh1), 5´→3´ exonuclease (Xrn1), exosome (Rrp6, Csl4, Rrp40, Rrp4, and Rrp44), TRAMP (Air1, Trf5, Mtr4, Air2, and Trf4), Ski (Ski2, Ski3, Ski7, and Ski8), and NMD (Upf1, Upf2, Upf3, and Nmd4). The molecular weight including the weight of the TAP tag (in kDa) is indicated for each factor. The band at ~25 kDa is caused by cross-reactivity of the light chain of the used antibodies for IP and Western Blot. A shift to higher molecular weight than indicated can be caused by UV-crosslinking of proteins to RNA.

DOI: https://doi.org/10.7554/eLife.47040.005

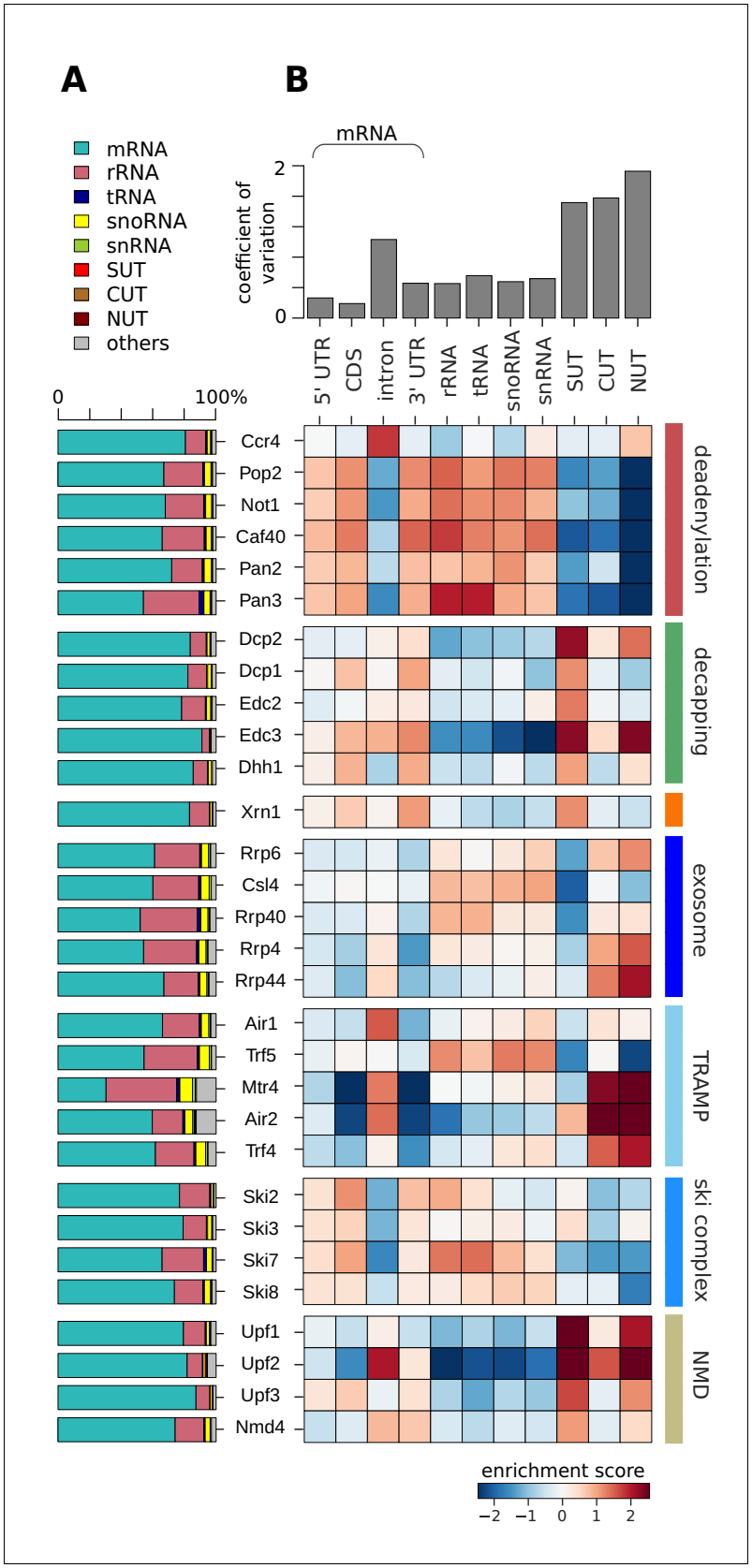


Figure 2. Distribution of degradation factor cross-link sites over the yeast transcriptome. (**A**) Fractions of high confidence PAR-CLIP sequencing reads of 30 yeast degradation factors fall into various transcript classes.

Figure 2 continued on next page

Figure 2 continued

Depicted classes are the following: messenger RNA (mRNA) in turquoise (n = 4,928), ribosomal RNA (rRNA) in antique pink (n = 24), transfer RNA (tRNA) in dark blue (n = 299), small nucleolar RNA (snoRNA) in yellow (n = 77), small nuclear RNA (snRNA) in green (n = 6), stable unannotated transcripts (SUTs) in red (n = 318), cryptic unstable transcripts (CUTs) in light brown (n = 637), Nrd1-unterminated transcripts (NUTs) in dark brown (n = 298)(Materials and methods). (**B**) Enrichment *z*-scores of high confidence PAR-CLIP cross-link sites of 30 yeast degradation factors (rows) in various segments of mRNA transcripts (left columns; UTR: untranslated region; intron; CDS: coding sequence), or other transcript classes as in A (other columns). The color-coded enrichment score shows the column and row normalized enrichment values of binding preferences of each factor for each transcript class (color encoded, depleted in blue and enriched in red). The coefficient of variation on top is the standard deviation divided by the mean for each transcript class.

DOI: https://doi.org/10.7554/eLife.47040.006

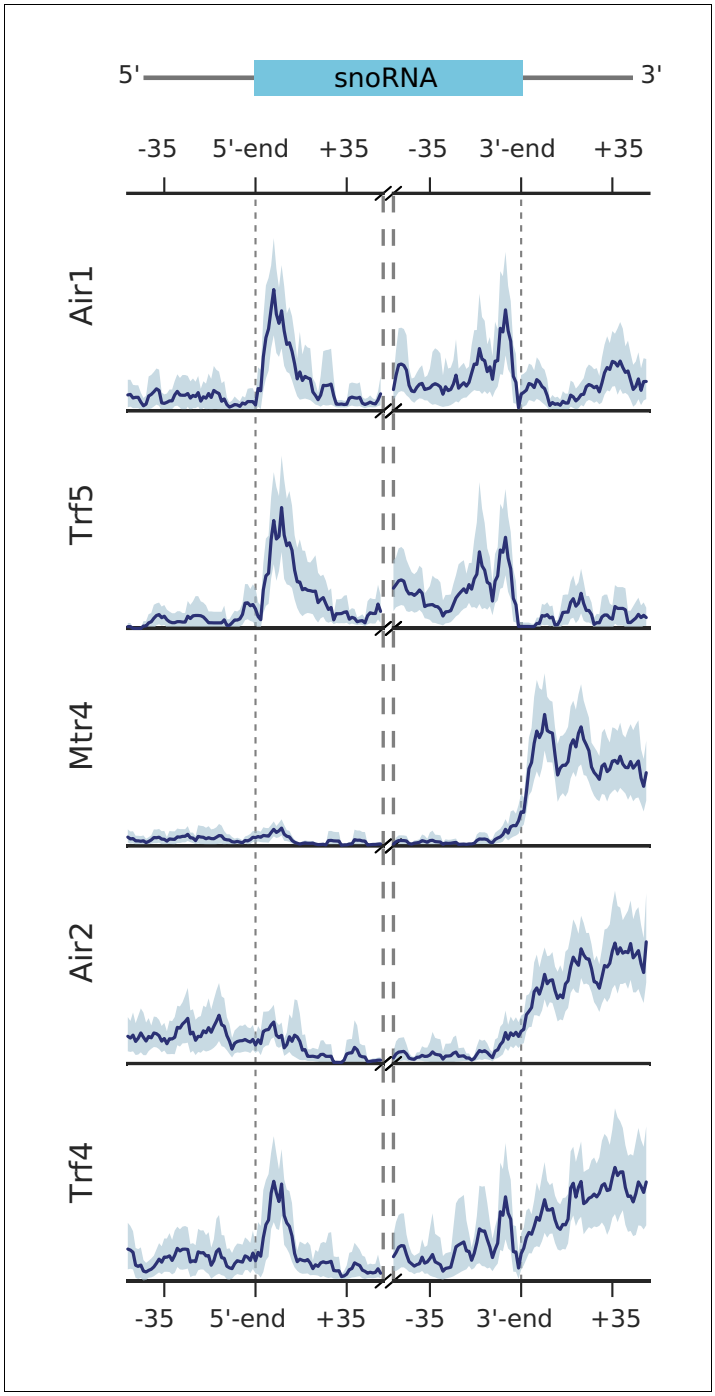


Figure 2—figure supplement 1. Metagene profiles for subunits of the TRAMP complexes on snoRNA genes. Transcript averaged PAR-CLIP occupancy profiles are shown for Air1, Trf5, Mtr4, Air2, and Trf4. snoRNA genes are aligned either at their 5′ end or at their 3′ end (n = 77). Occupancy profiles are shown over the range of ±50 nt.
DOI: https://doi.org/10.7554/eLife.47040.007

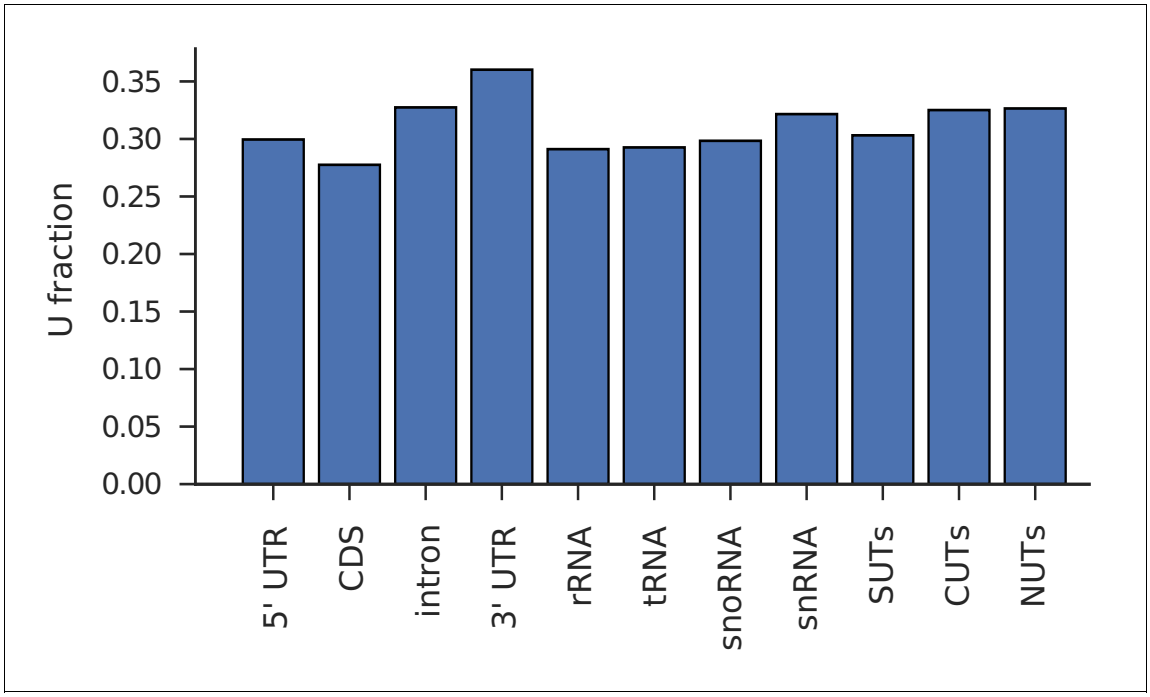


Figure 2—figure supplement 2. Different transcript classes have comparable U-content. Fraction of U over all bases in transcript classes studied in ***Figure 2*** (untranslated region (UTR); intron; coding sequence (CDS), ribosomal RNA (rRNA), transfer RNA (tRNA), small nucleolar RNA (snoRNA), small nuclear RNA (snRNA), stable unannotated transcripts (SUTs), cryptic unstable transcripts (CUTs), Nrd1- unterminated transcripts (NUTs)).
DOI: https://doi.org/10.7554/eLife.47040.008

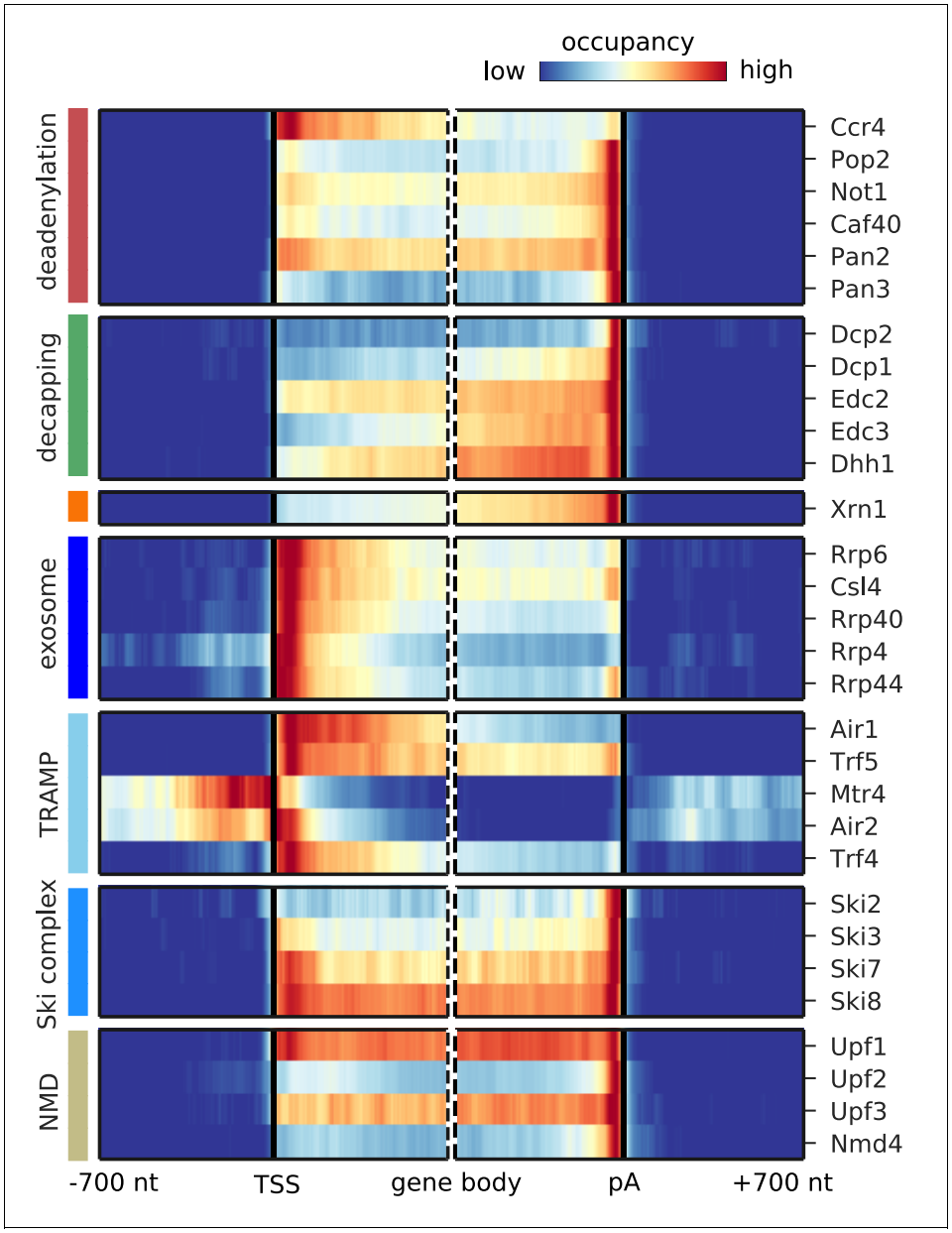


Figure 3. Metagene analysis of degradation factor binding on mRNAs. Averaged occupancy profiles of degradation factors over mRNAs aligned around their transcription start site (TSS) (n = 3,193, left) and around their poly-adenylation (pA) site (n = 3,193, right) in a window of [±700 nt]. Regions that have neighboring transcripts on the same strand were removed to avoid contaminating profiles (Materials and methods). Factors are grouped according to their functional role; from top to bottom: deadenylation, decapping, Xrn1, exosome, TRAMP complex, Ski complex, or NMD. The color code shows the average occupancy normalized between the minimum (blue) and maximum (red) values per profile.

DOI: https://doi.org/10.7554/eLife.47040.009

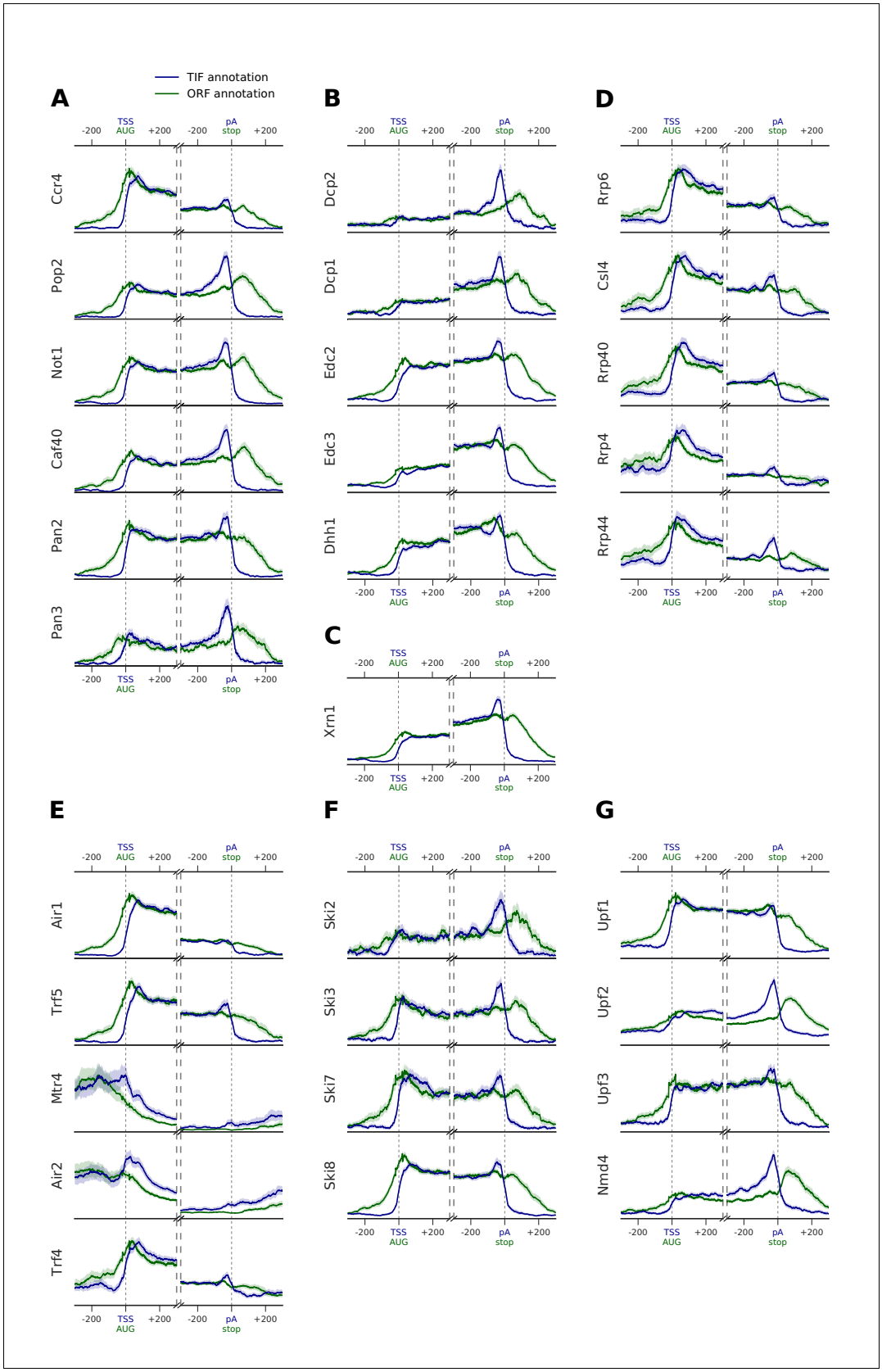


Figure 3—figure supplement 1. Metagene profiles of yeast RNA degradation factors centered on translation start and stop sites in comparison to TIF-annotated TSS and pA sites. Transcript-averaged PAR-CLIP occupancy profiles are shown for RNA degradation factors involved in (**A**) deadenylation, (**B**)

Figure 3—figure supplement 1 continued on next page

Figure 3—figure supplement 1 continued

decapping, (**C**) 5′→3′ exonuclease Xrn1, (**D**) exosome, (**E**) TRAMP, (**F**) Ski, and (**G**) NMD. Transcripts are aligned either at transcript start site (TSS) and poly-adenylation (pA) site (marked with blue) or at their start and stop codons (marked with green). TIF-seq based annotation is shown in blue (n = 3193 for TSS and pA site profiles) (***Pelechano et al., 2013***). Open reading frames (ORF) annotated in the SGD (version 64.2.1) are shown in green (n = 4012 for TSS, and n = 3965 for pA site selected transcripts). To avoid contaminating signals from neighboring genes, we filtered out regions that had annotations upstream and downstream of the centered gene (up to 700 nt) (Materials and methods). Shaded areas (in blue TIF-seq annotation, or in green for ORF annotation) depict 95% confidence intervals derived from bootstrapping genes. Comparison between these two profiles highlights preferences for end binding degradation factors in binding to untranslated regions at the two sides of the transcript.

DOI: https://doi.org/10.7554/eLife.47040.010

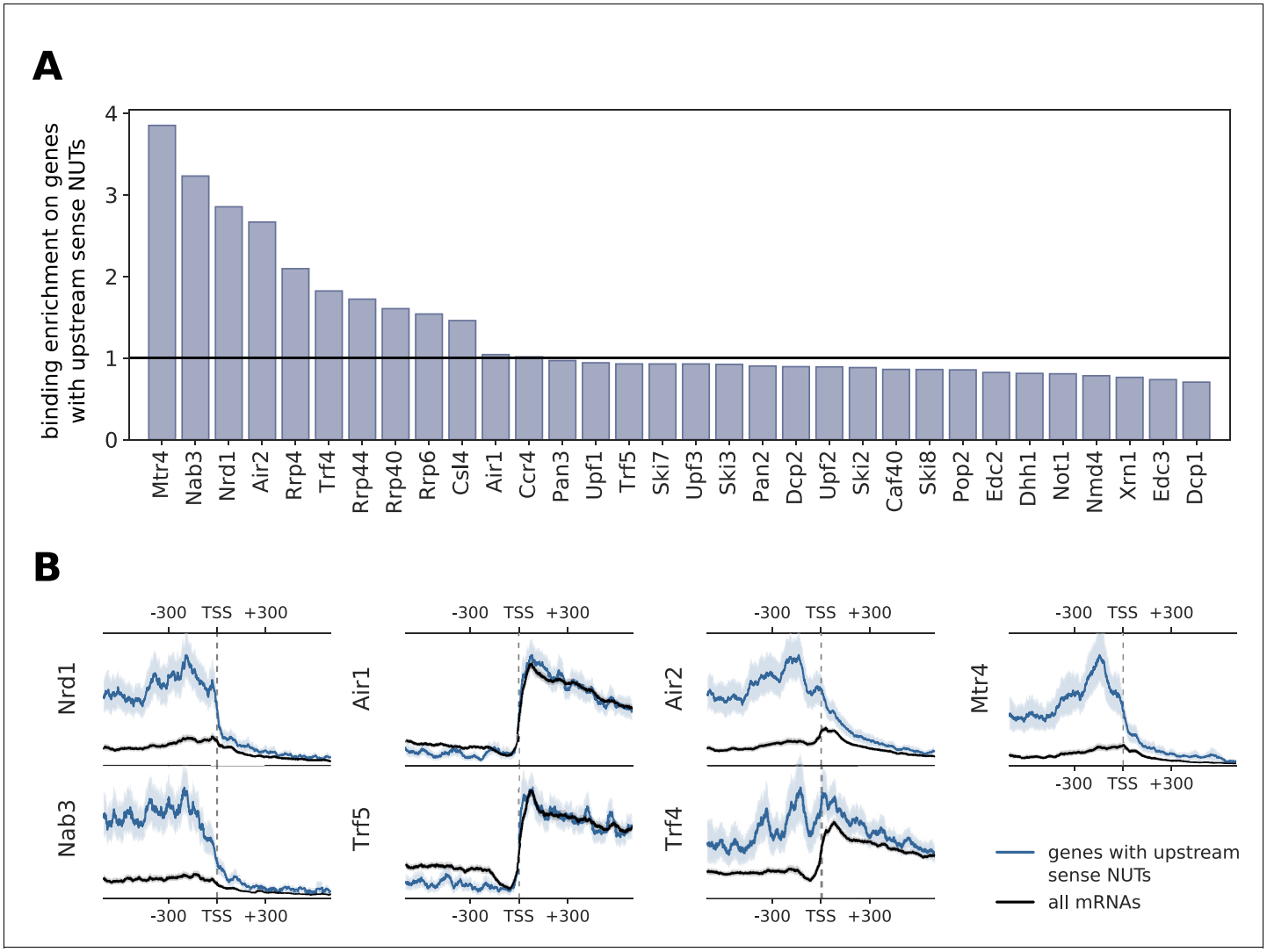


Figure 3—figure supplement 2. Comparison of binding profiles on genes containing annotated upstream sense NUTs with all mRNAs. (**A**) Binding enrichment of degradation factors around the TSS of genes with an upstream sense NUT. Enrichment is defined as the ratio of the average occupancy in the interval [±300 nt] of the TSS on these genes that contain an upstream NUT (n = 459) (***Schulz et al., 2013***) divided by the average occupancy on all genes. (**B**) Transcript-averaged PAR-CLIP occupancy profiles for all mRNAs (black) is compared to patterns derived from genes with upstream sense NUTs (blue). Transcripts were aligned at their TSS and averaged over the interval of [±600 nt]. We compared Nrd1 and Nab3 profiles, known to process NUTs, with subunits of the TRAMP complex. 95% confidence intervals obtained from bootstrapping genes are shown with gray and blue shades.
DOI: https://doi.org/10.7554/eLife.47040.011

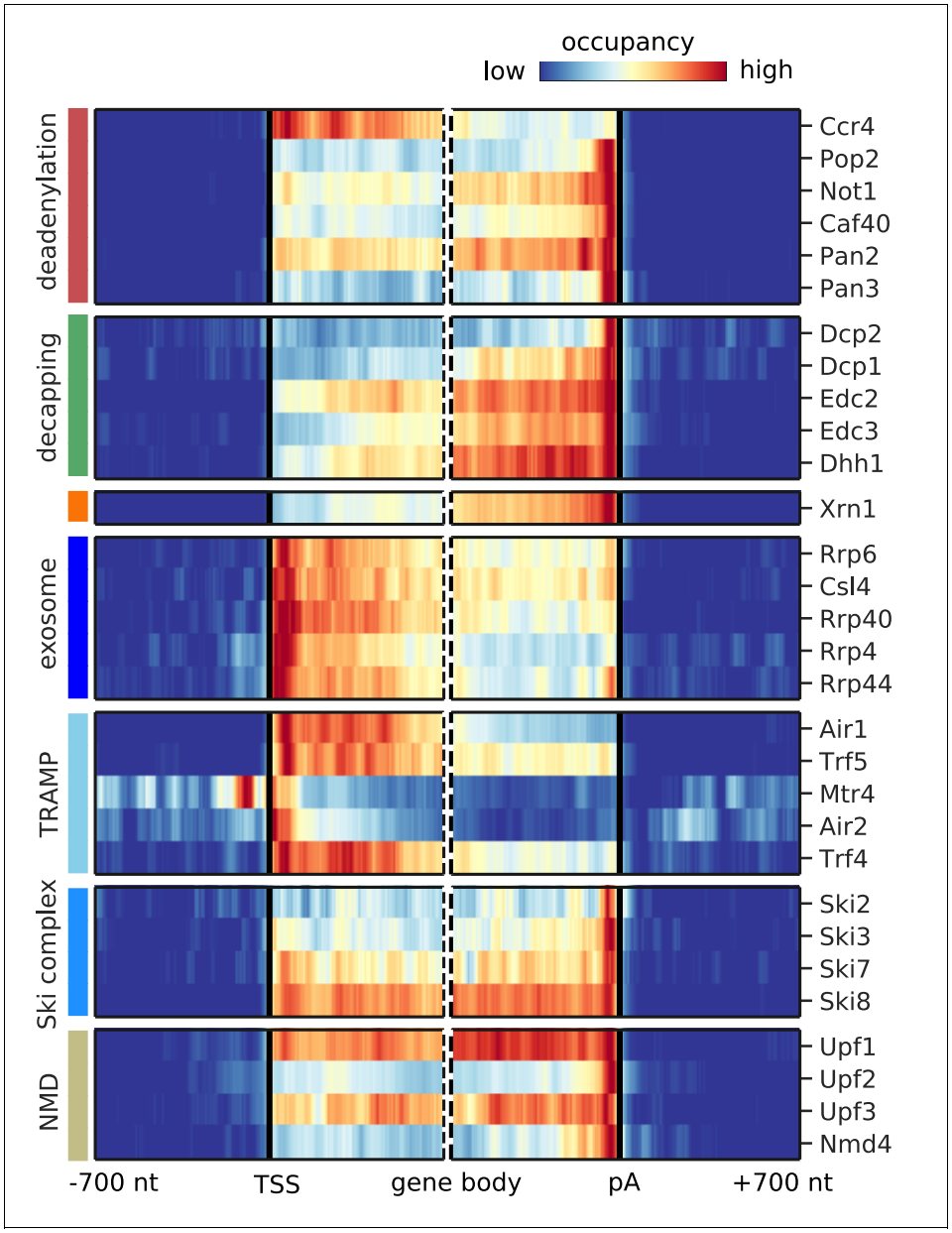


Figure 3—figure supplement 3. Metagene analysis of degradation factor binding on mRNAs after removing signals from known NUTs and CUTs. Cross-link sites were filtered to exclude regions that were previously annotated as NUTs and CUTs (***Neil et al., 2009***; ***Schulz et al., 2013***). Averaged occupancy profiles of degradation factors are then shown over mRNAs aligned around their transcription start site (TSS) (n = 3,193, left) and around their poly-adenylation (pA) site (n = 3,193, right) in a window of [±700 nt]. Regions that have neighboring transcripts on the same strand were removed to avoid contaminating profiles (Materials and methods). Factors are grouped according to their functional role; from top to bottom: deadenylation, decapping, Xrn1, exosome, TRAMP complex, Ski complex, or NMD. The color code shows the average occupancy normalized between the minimum (blue) and maximum (red) values per profile.

DOI: https://doi.org/10.7554/eLife.47040.012

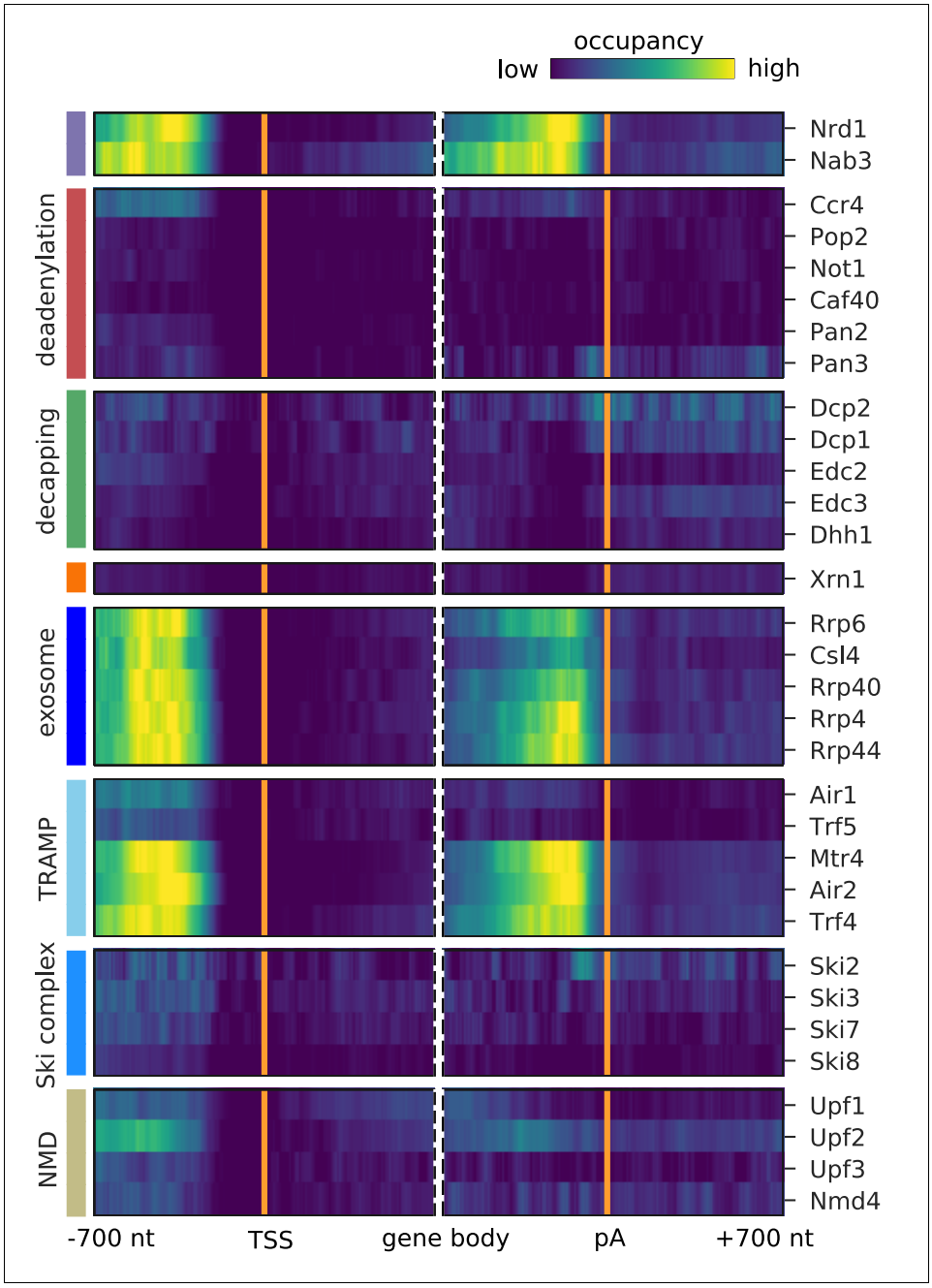


Figure 4. Surveillance of aberrant nuclear antisense RNAs by the exosome and the TRAMP4 complex. Averaged occupancy profiles of degradation factors binding to transcripts antisense of mRNAs aligned around transcription start site (TSS) (n = 3,076, left) and around their poly-adenylation (pA) site (n = 2,705, right) in a window of [±700 nt]. Regions with annotated genes on the antisense strand are removed to avoid contaminating the profiles (Materials and methods). The color code shows the average occupancy normalized between the minimum (blue) and maximum (yellow) values per profile. On top, previously published PAR-CLIP profiles for Nrd1 and Nab3 are included for comparison (***Schulz et al., 2013***).

DOI: https://doi.org/10.7554/eLife.47040.013

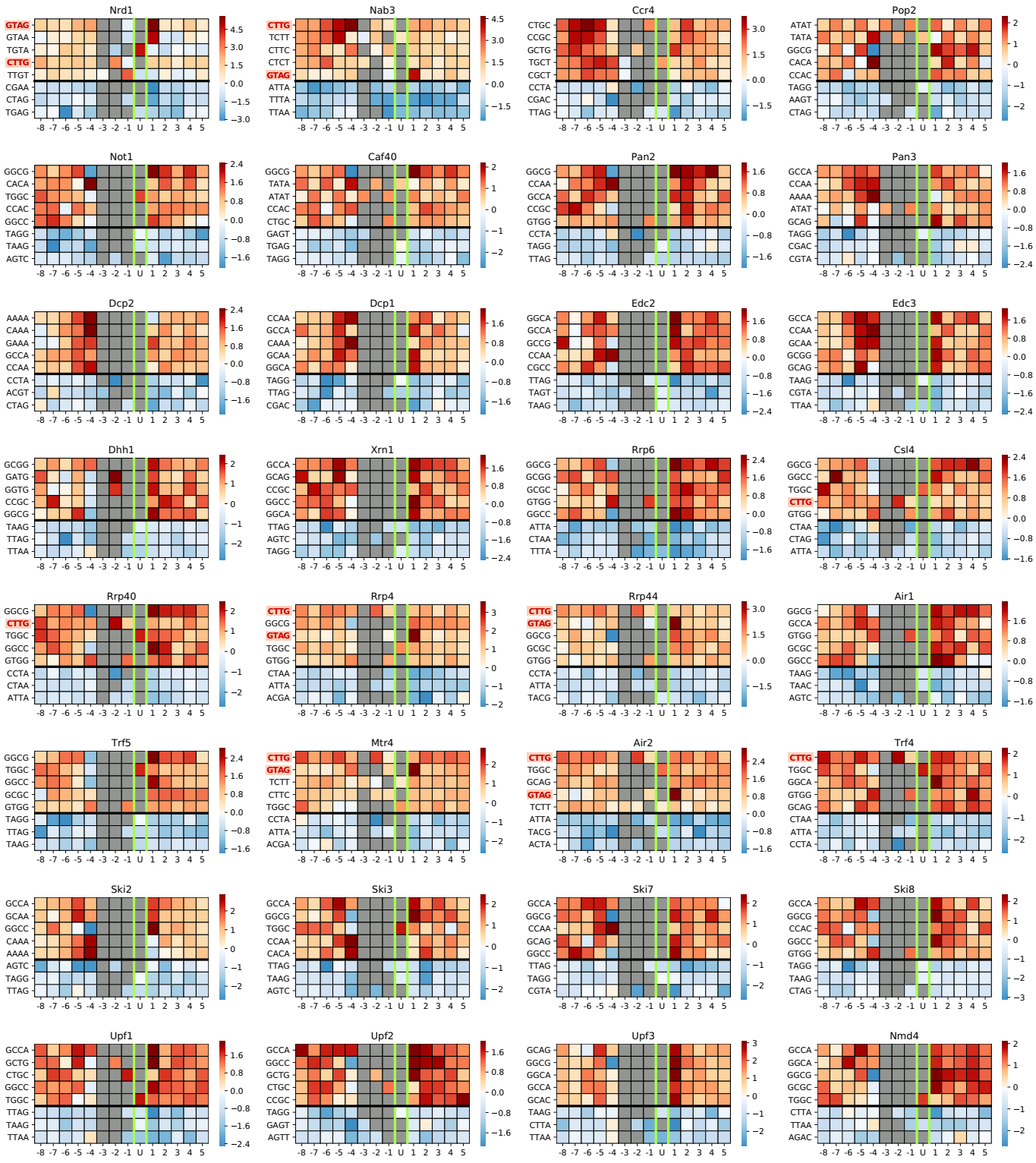


Figure 4—figure supplement 1. Motif enrichment analysis shows enrichment of Nrd1/Nab3 motifs for the TRAMP4 and the exosome complex. Motif analysis was performed for all degradation factors in this study. Nrd1 and Nab3 are included for comparison (***Schulz et al., 2013***). Occurrences of Nrd1 motif (GTAG) and Nab3 motif (CTTG) are highlighted with red. The color code shows the log2 enrichment factor of top five enriched and top 3

Figure 4—figure supplement 1 continued on next page

Figure 4—figure supplement 1 continued

depleted 4-mers around PAR-CLIP cross-link sites [±8 nt]. Dark red represents strong enrichment and dark blue shows strong depletion of a 4-mer. Infeasible combinations are shown with gray.

DOI: https://doi.org/10.7554/eLife.47040.014

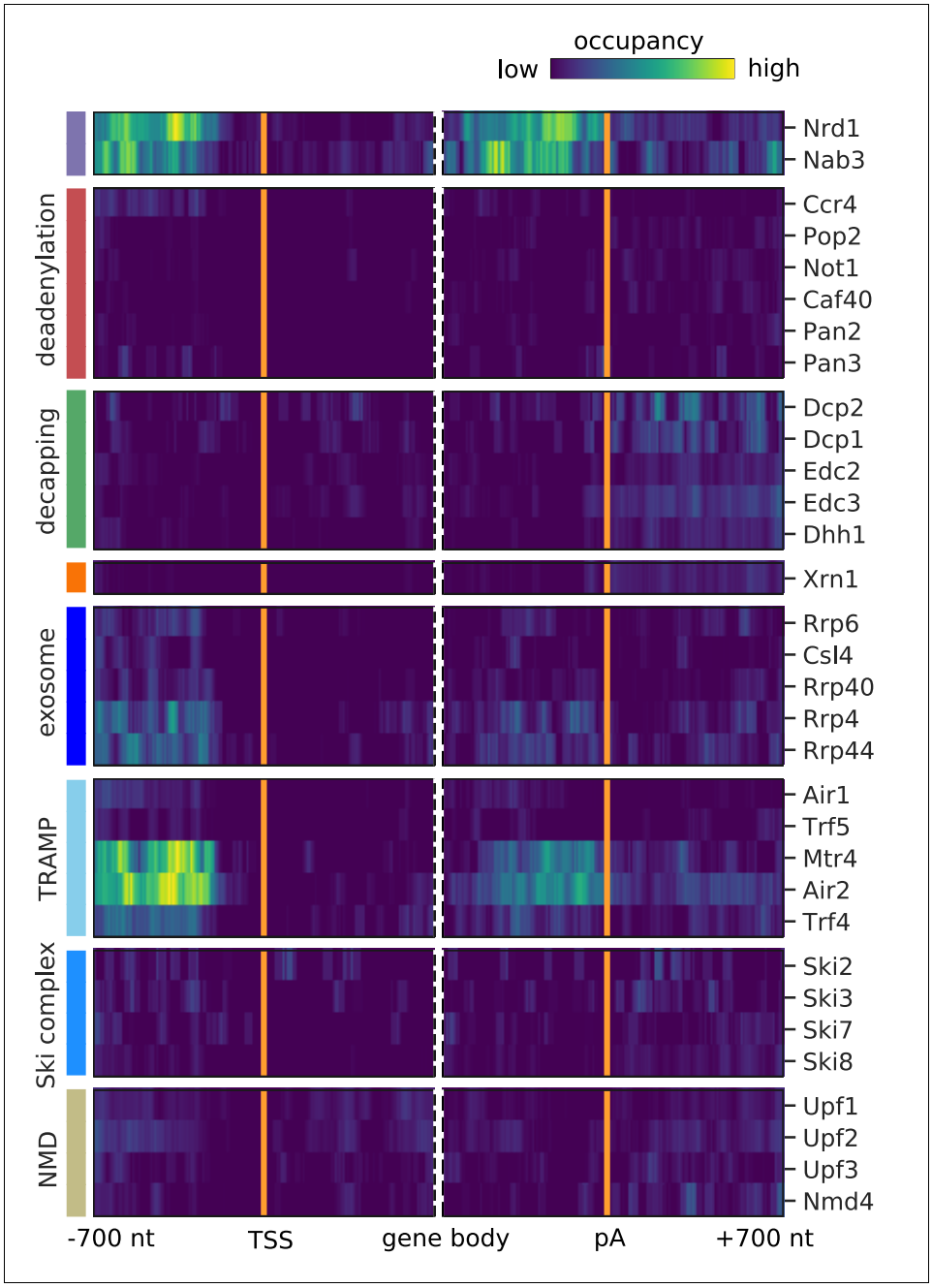


Figure 4—figure supplement 2. The aberrant nuclear ncRNAs bound by components of the exosome and the TRAMP4 complex are primarily NUTs and CUTs. Cross-link sites were filtered to exclude regions that were previously annotated as NUTs and CUTs (***Neil et al., 2009***; ***Schulz et al., 2013***). Averaged occupancy profiles of degradation factors are shown on transcripts antisense of mRNAs aligned around the transcription start site (TSS) (n = 3,076, left) and around their poly-adenylation (pA) site (n = 2,705, right) in a window of [±700 nt]. Regions with annotated genes on the antisense strand are removed to avoid contaminating the profiles (Materials and methods). The color code shows the average occupancy normalized between the minimum (blue) and maximum (yellow) values per profile. On top, previously published PAR-CLIP profiles for Nrd1 and Nab3 are included for comparison (***Schulz et al., 2013***).

DOI: https://doi.org/10.7554/eLife.47040.015

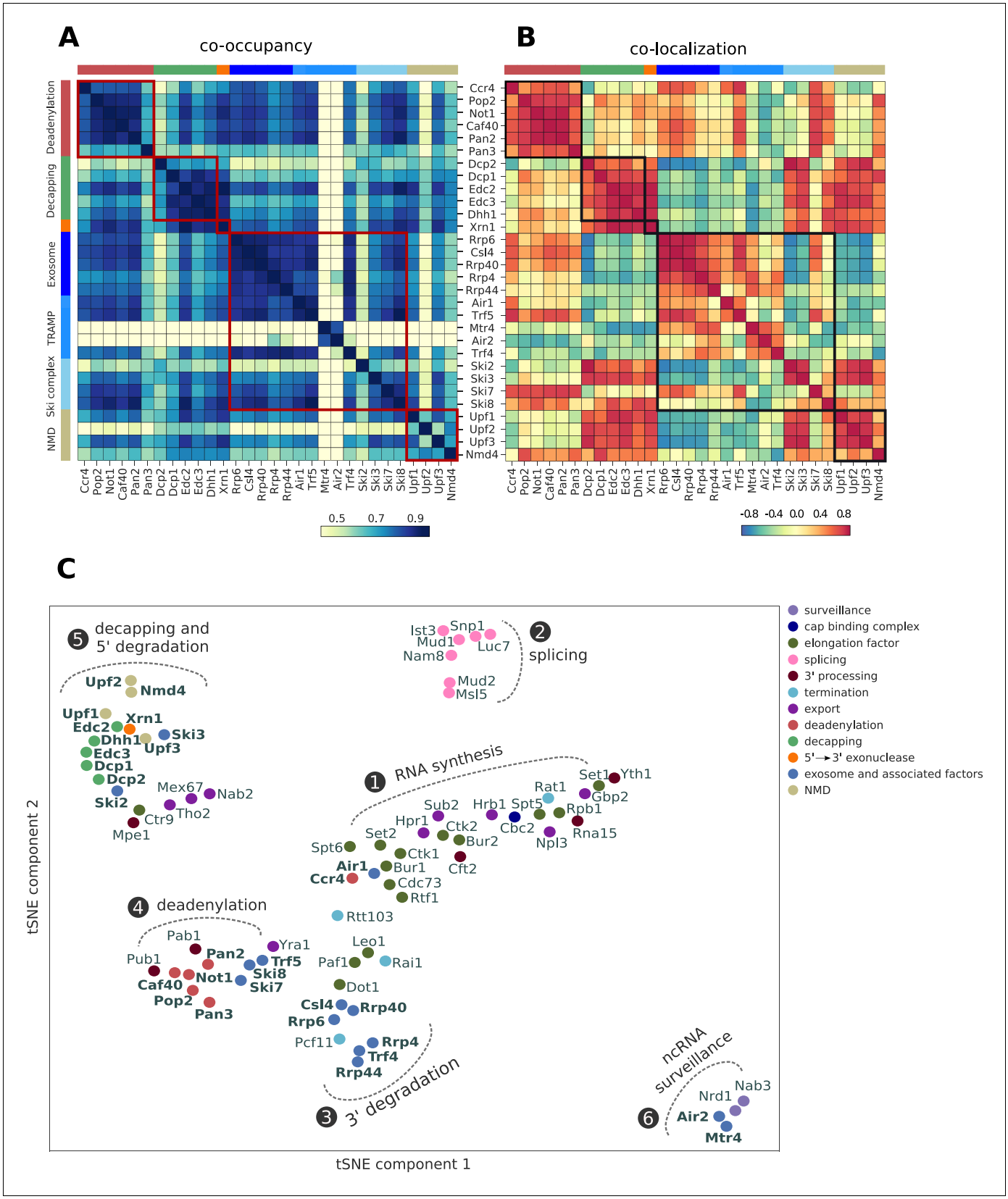


Figure 5. Global co-occupancy and co-localization analysis reveals unexpected cooperation between factors from different complexes and pathways. (**A**) Matrix of pairwise correlation coefficients of factor occupancies evaluated over all transcripts. (**B**) Matrix of co-localization based on the enrichment

Figure 5 continued on next page

Figure 5 continued

of factor x binding within 40 nt of the cross-link site of factor x´ (Materials and methods). (**C**) Two-dimensional embedding of the co-occupancies in (**A**) analyzed for 74 RNA processing factors with tSNE, including 30 factors from this study (highlighted in bold), and 44 factors from previous studies (***Baejen et al., 2017***; ***Baejen et al., 2014***; ***Battaglia et al., 2017***; ***Schulz et al., 2013***) (***Supplementary file 1***). Factors that are plotted in close proximity show a preference for binding to the same transcripts. Clusters present factors involved in RNA synthesis (1), splicing (2), 3´ processing (3), deadenylation (4), decapping (5), nuclear ncRNA processing (6), and surveillance (7).

DOI: https://doi.org/10.7554/eLife.47040.016

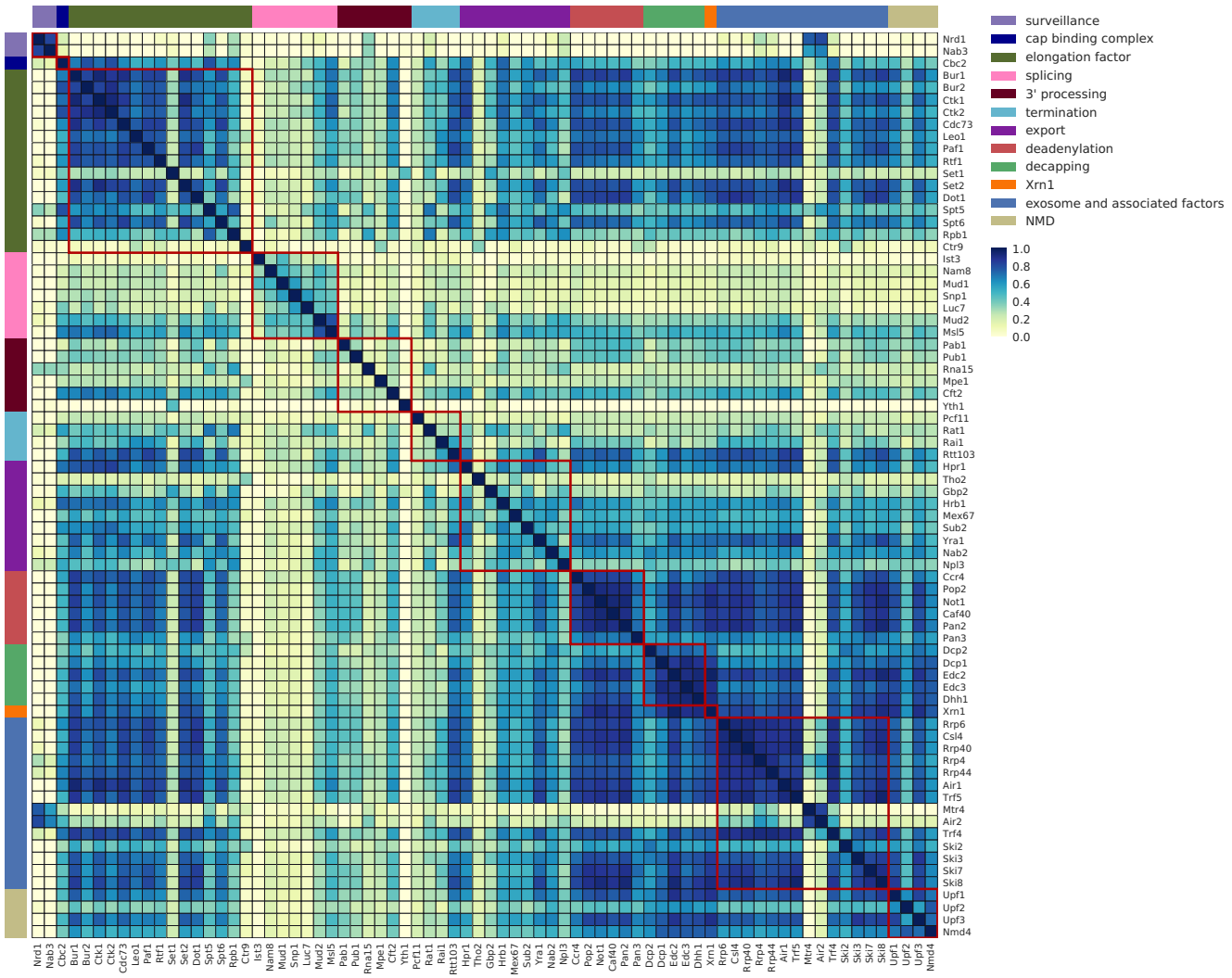


Figure 5—figure supplement 1. Co-occupancy for 74 RNA processing factors. Matrix of pairwise correlation coefficients of factor occupancies evaluated over all transcripts. Analysis for 74 RNA processing factors, including 30 factors from this study, and 44 factors from previous studies (***Baejen et al., 2017***; ***Baejen et al., 2014***; ***Battaglia et al., 2017***; ***Schulz et al., 2013***) (see ***Supplementary file 1***). Dark blue represents high correlation in binding across all transcripts. Factors are sorted and color coded (left and upper border) according to their general function.
DOI: https://doi.org/10.7554/eLife.47040.017

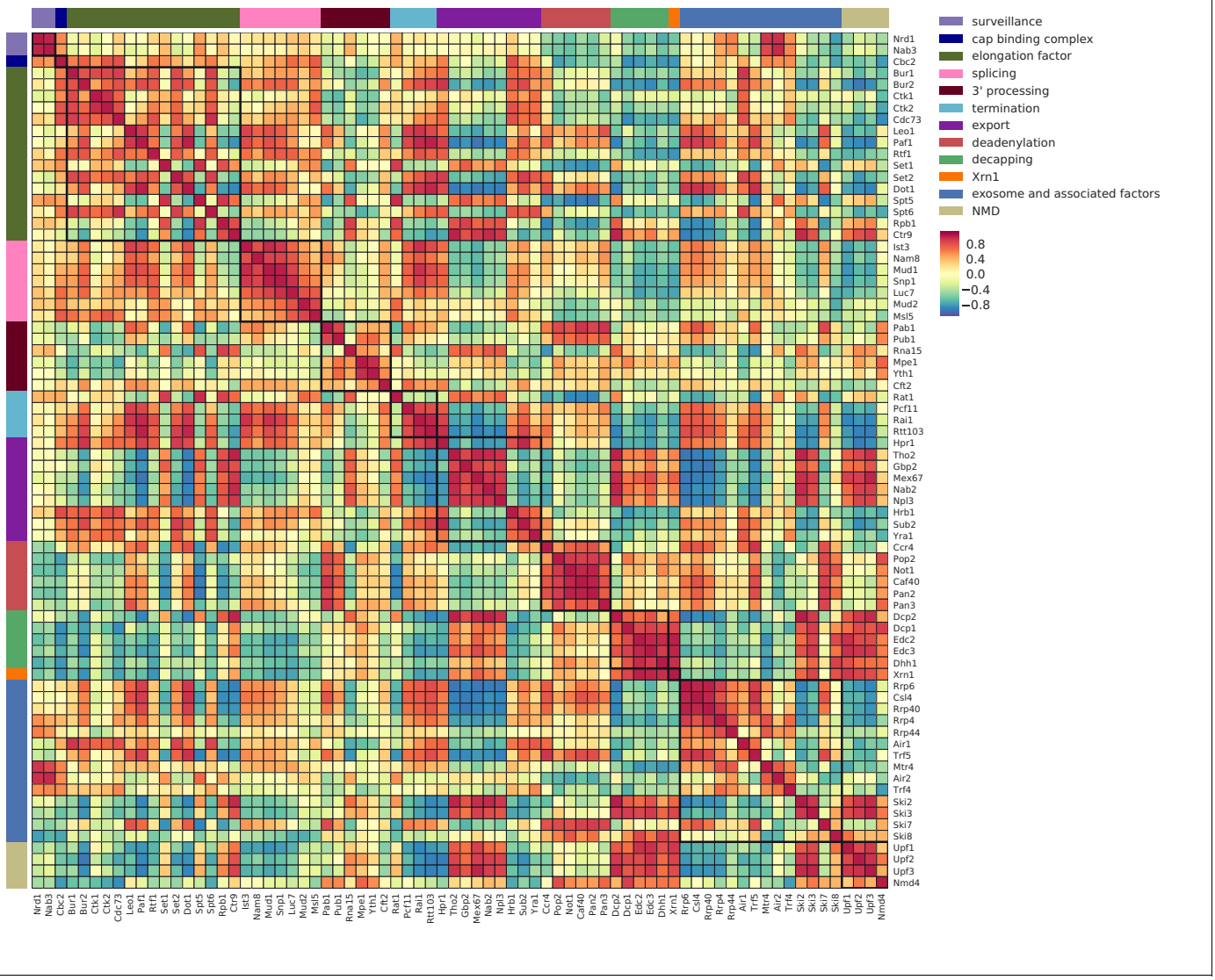


Figure 5—figure supplement 2. Co-localization coefficients for all 74 RNA processing factors. Pairwise correlation between normalized co-localization profiles of factors in a window of 40 nt centered at PAR-CLIP cross-link sites. Analysis for 74 RNA processing factors, including 30 factors from this study, and 44 factors from previous studies (***Baejen et al., 2017***; ***Baejen et al., 2014***; ***Battaglia et al., 2017***; ***Schulz et al., 2013***) (see ***Supplementary file 1***). High co-localization represents binding to the same position on transcripts (marked with dark red). Factors are sorted and color coded (left and upper border) according to their general function.

DOI: https://doi.org/10.7554/eLife.47040.018

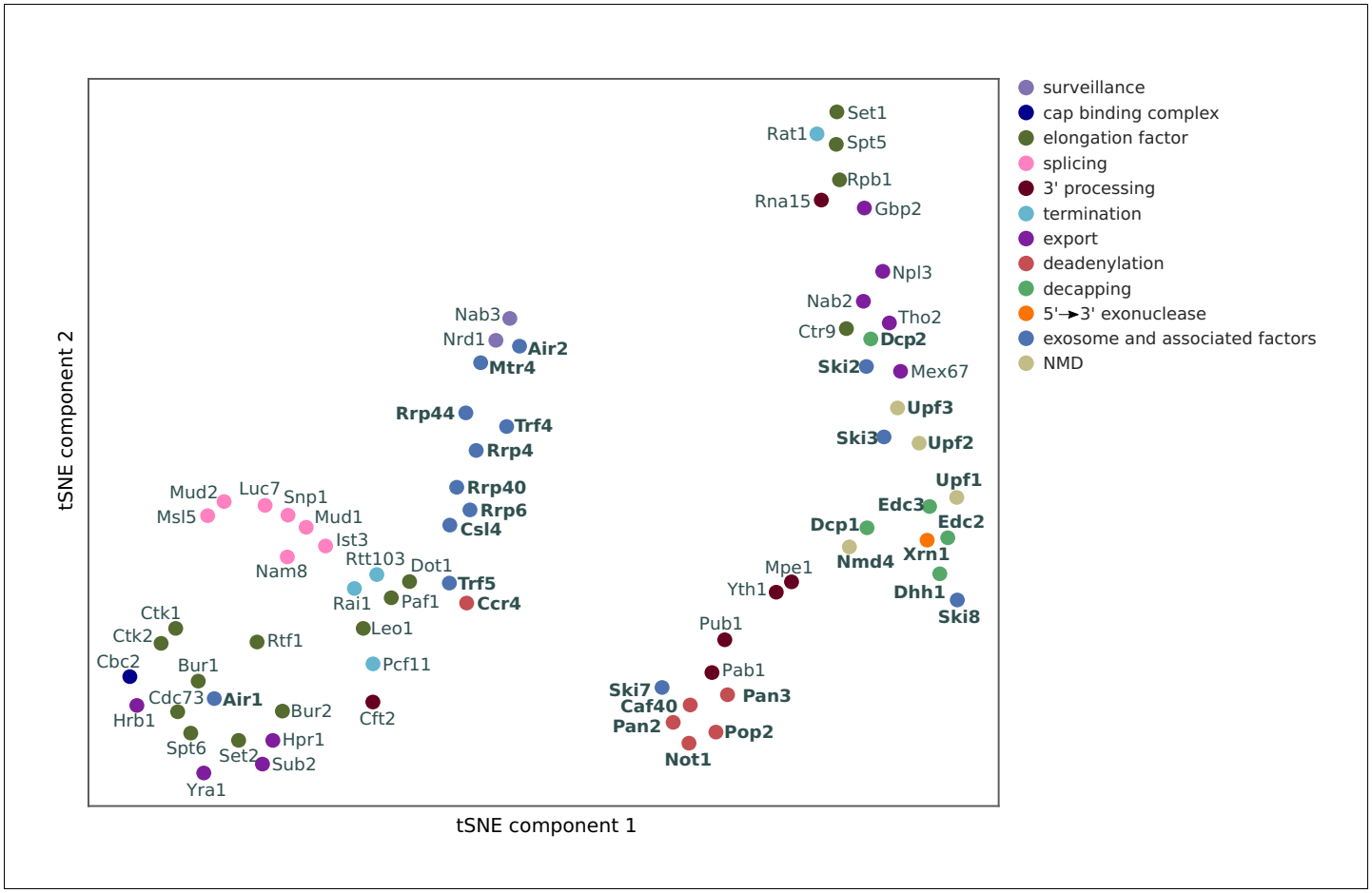


Figure 5—figure supplement 3. Two-dimensional embedding of co-localization between 74 RNA processing factors. tSNE plot visualizes similarities in co-localization profiles between RNA processing factors (factors are color-coded based on their function). Analysis for 74 RNA processing factors, including 30 factors from this study (marked in bold), and 44 factors from previous studies (***Baejen et al., 2017***; ***Baejen et al., 2014***; ***Battaglia et al., 2017***; ***Schulz et al., 2013***) (see ***Supplementary file 1***).

DOI: https://doi.org/10.7554/eLife.47040.019

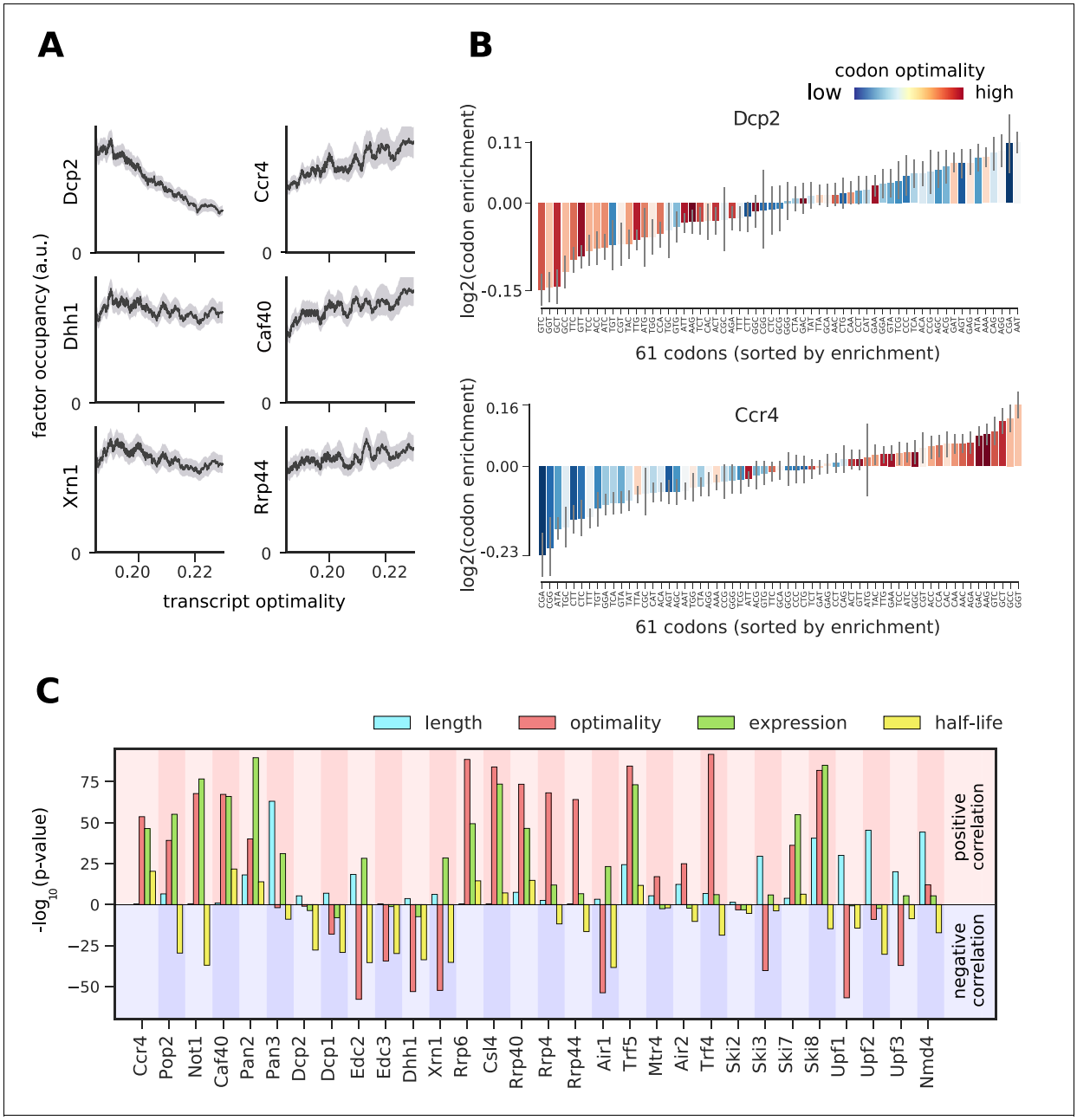


Figure 6. Binding preferences reveal a link between decapping-mediated degradation and translation. (**A**) Total occupancy per mRNA (according to TIF-seq annotation) for six factors as a function of the average mRNA codon optimality (transcript optimality). The occupancy of factors from the 5′→3′ degradation machinery (decapping and Xrn1, left) decreases with increasing transcript optimality, whereas the occupancy of factors from the 3′→5′ degradation machinery (Ccr4 and Caf40 deadenylation complex subunits and exosome subunit Rrp44, right) increases with increasing average codon optimality. (Gray shading: 95% confidence intervals generated by bootstrapping mRNAs). (**B**) Codon enrichment in transcripts bound by Dcp2 and Ccr4 compared to the average frequency over all mRNAs. The bar colors represent codon optimality, with highly optimal codons shown in dark red. (Thin gray lines: 90% confidence intervals generated by bootstrapping coding sequences.) (**C**) Significance of correlations between the binding strength of degradation factors and transcript length, transcript optimality (***Pechmann and Frydman, 2013***), expression level (***Baejen et al., 2017***), and half-life derived by multivariate linear regression analysis (Materials and methods). Bars are separated according to the direction of correlation with positive correlation marked by a red background and negative correlation marked by a blue background.

DOI: https://doi.org/10.7554/eLife.47040.020

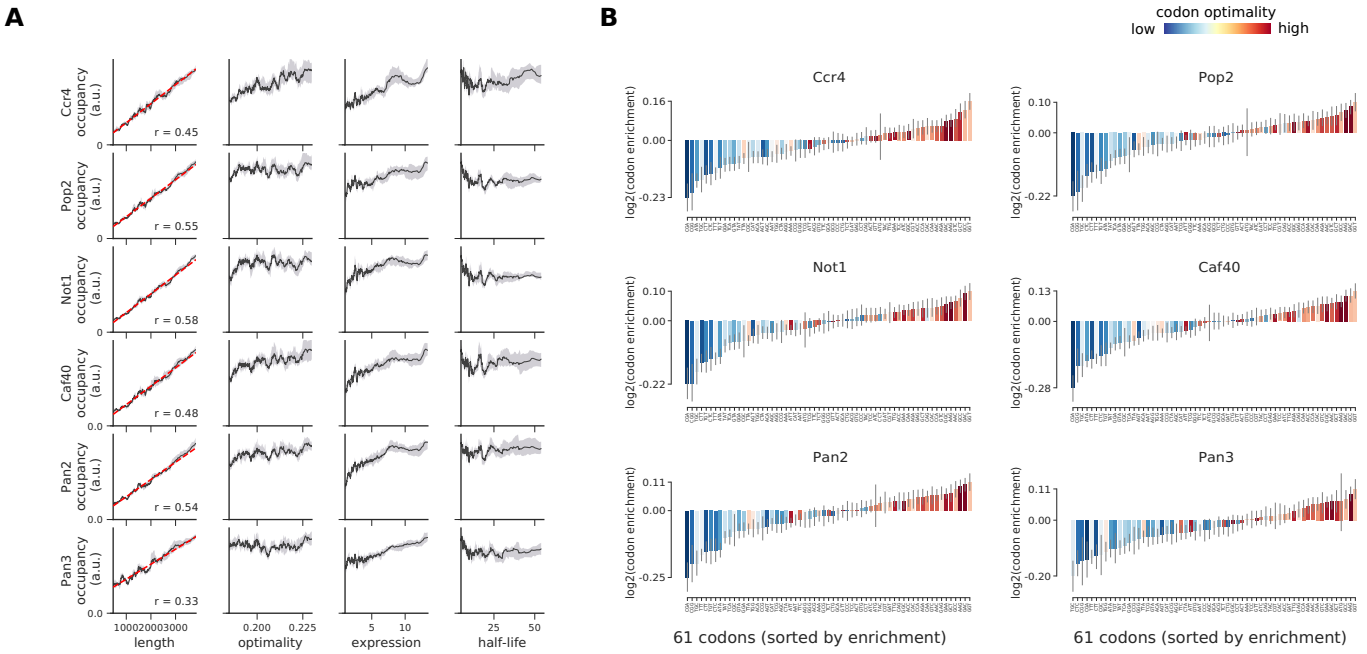


Figure 6—figure supplement 1. Occupancies of deadenylation factors (Ccr4, Pop2, Not1, Caf40, Pan2, and Pan3) compared to transcript length, optimality, expression level, and half-life. (**A**) To understand binding specificity of deadenylation factors, the total occupancy of each factor on a transcript is plotted against various transcript features (Gray shading: 95% confidence intervals generated by bootstrapping transcripts). (**B**) Same analysis as in ***Figure 6B***: Codon enrichment shows deviations in codon frequencies of transcripts bound by a degradation factor compared to each codon's frequency on all coding sequences. Each bar is colored according to its codon-optimality with highly optimal codons in dark red and highly non-optimal codons in dark blue. (Gray lines: 90% confidence intervals generated by bootstrapping coding sequences).

DOI: https://doi.org/10.7554/eLife.47040.021

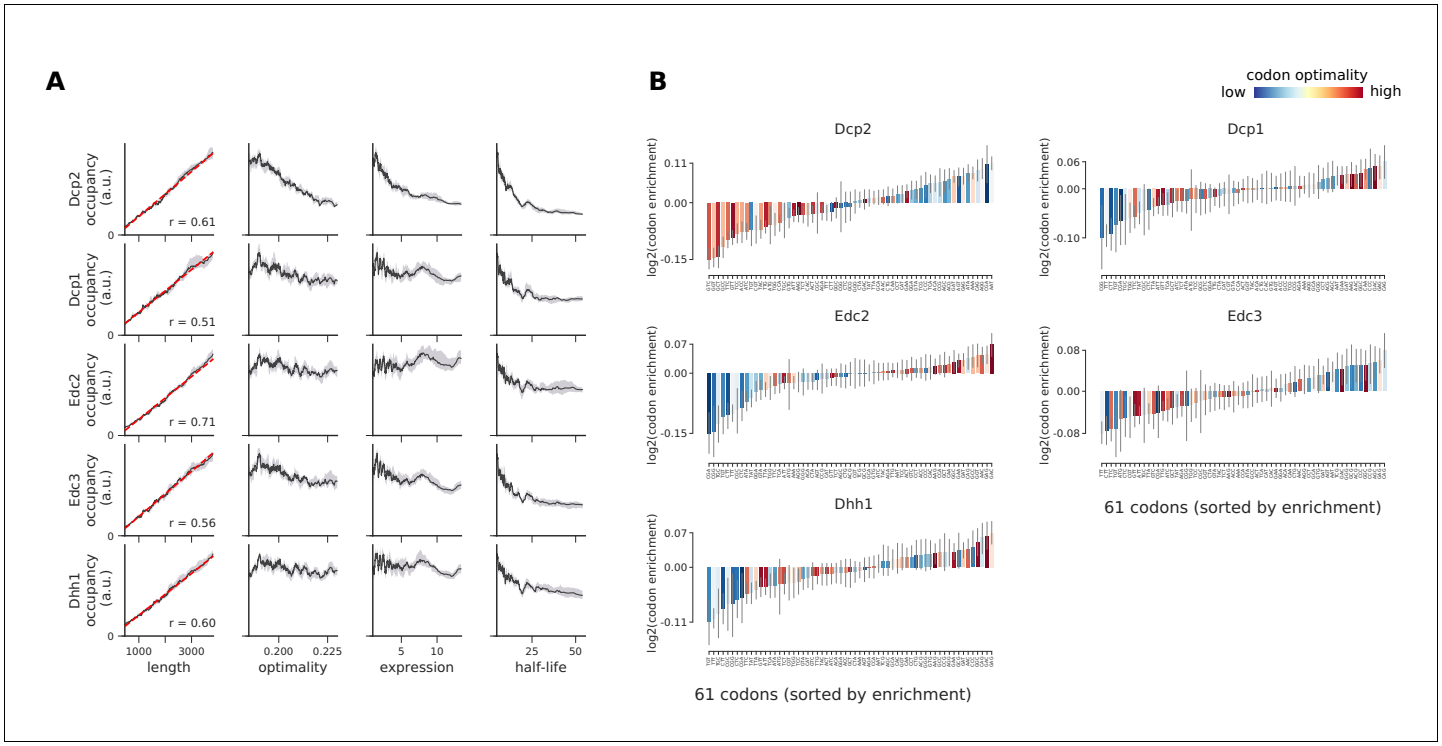


Figure 6—figure supplement 2. Occupancies of decapping factors (Dcp2, Dcp1, Edc2, Edc3, and Dhh1) compared to transcript length, optimality, expression level, and half-life. (**A**) To understand binding specificity of decapping factors, the total occupancy of each factor on a transcript is plotted against various transcript features (Gray shading: 95% confidence intervals generated by bootstrapping transcripts). (**B**) Same analysis as in ***Figure 6B***: Codon enrichment shows deviations in codon frequencies of transcripts bound by a degradation factor compared to each codon's frequency on all coding sequences. Each bar is colored according to its codon-optimality with highly optimal codons in dark red and highly non-optimal codons in dark blue. (Gray lines: 90% confidence intervals generated by bootstrapping coding sequences).

DOI: https://doi.org/10.7554/eLife.47040.022

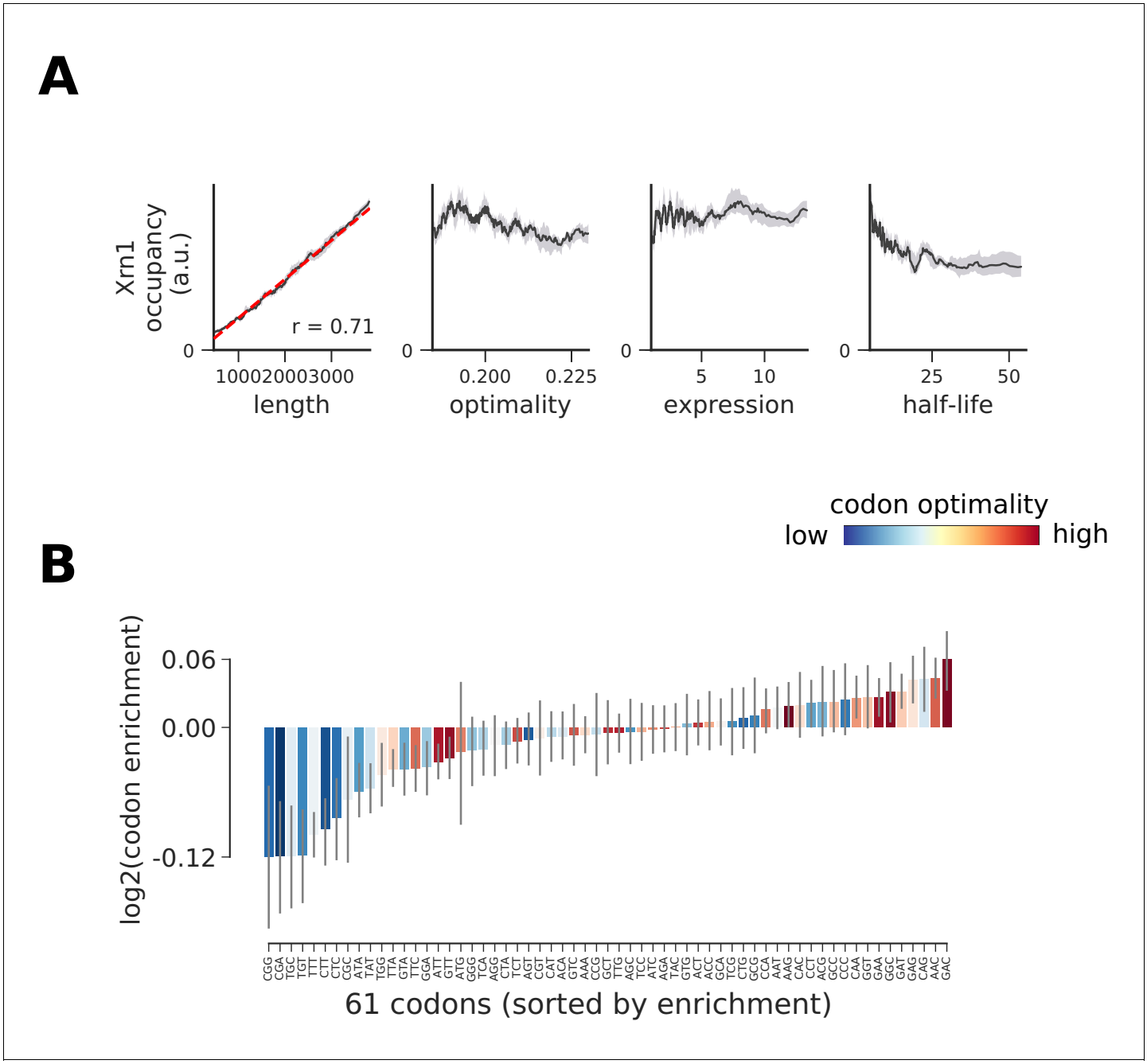


Figure 6—figure supplement 3. Occupancy of Xrn1 compared to transcript length, optimality, expression level, and half-life. (**A**) To understand binding specificity of Xrn1 on various mRNAs, the total occupancy of Xrn1 on a transcript is plotted against various transcript features (Gray shading: 95% confidence intervals generated by bootstrapping transcripts). (**B**) Same analysis as in ***Figure 6B***: Codon enrichment shows deviations in codon frequencies of transcripts bound by a degradation factor compared to each codon's frequency on all coding sequences. Each bar is colored according to its codon-optimality with highly optimal codons in dark red and highly non-optimal codons in dark blue. (Gray lines: 90% confidence intervals generated by bootstrapping coding sequences).

DOI: https://doi.org/10.7554/eLife.47040.023

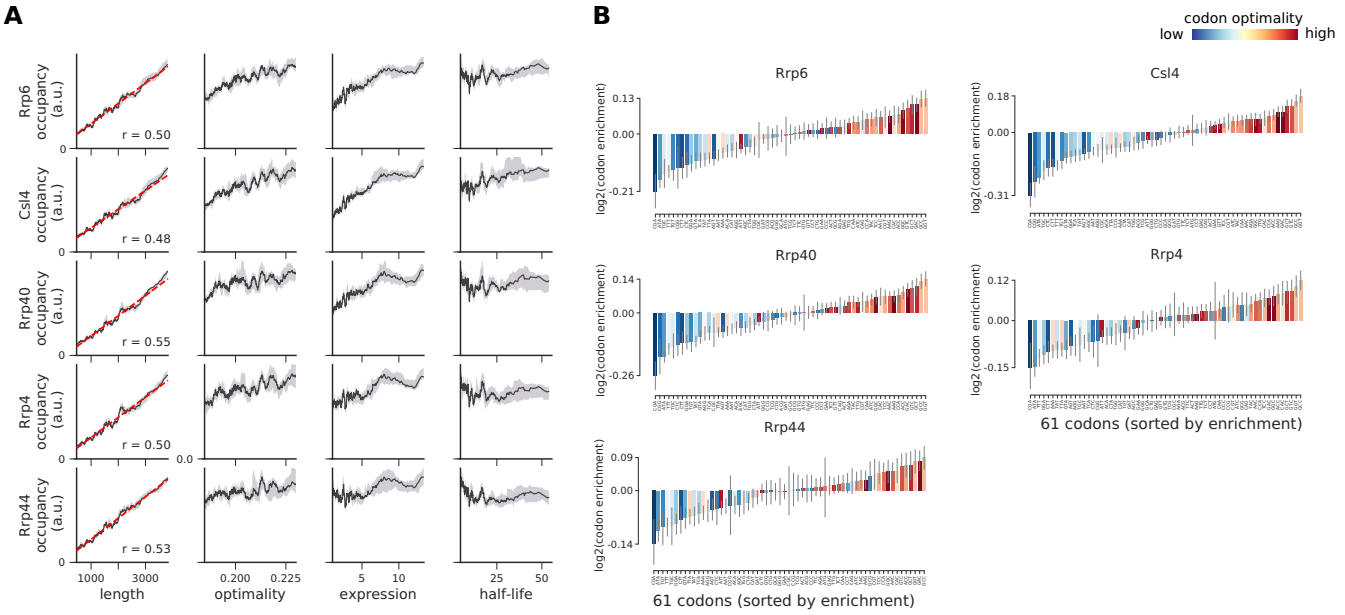


Figure 6—figure supplement 4. Occupancies of exosome components (Rrp6, Csl4, Rrp40, Rrp4, and Rrp44) compared to transcript length, optimality, expression level, and half-life. (**A**) To understand binding specificity of exosome components, the total occupancy of each factor on a transcript is plotted against various transcript features (Gray shading: 95% confidence intervals generated by bootstrapping transcripts). (**B**) Same analysis as in ***Figure 6B***: Codon enrichment shows deviations in codon frequencies of transcripts bound by a degradation factor compared to each codon's frequency on all coding sequences. Each bar is colored according to its codon-optimality with highly optimal codons in dark red and highly non-optimal codons in dark blue. (Gray lines: 90% confidence intervals generated by bootstrapping coding sequences).

DOI: https://doi.org/10.7554/eLife.47040.024

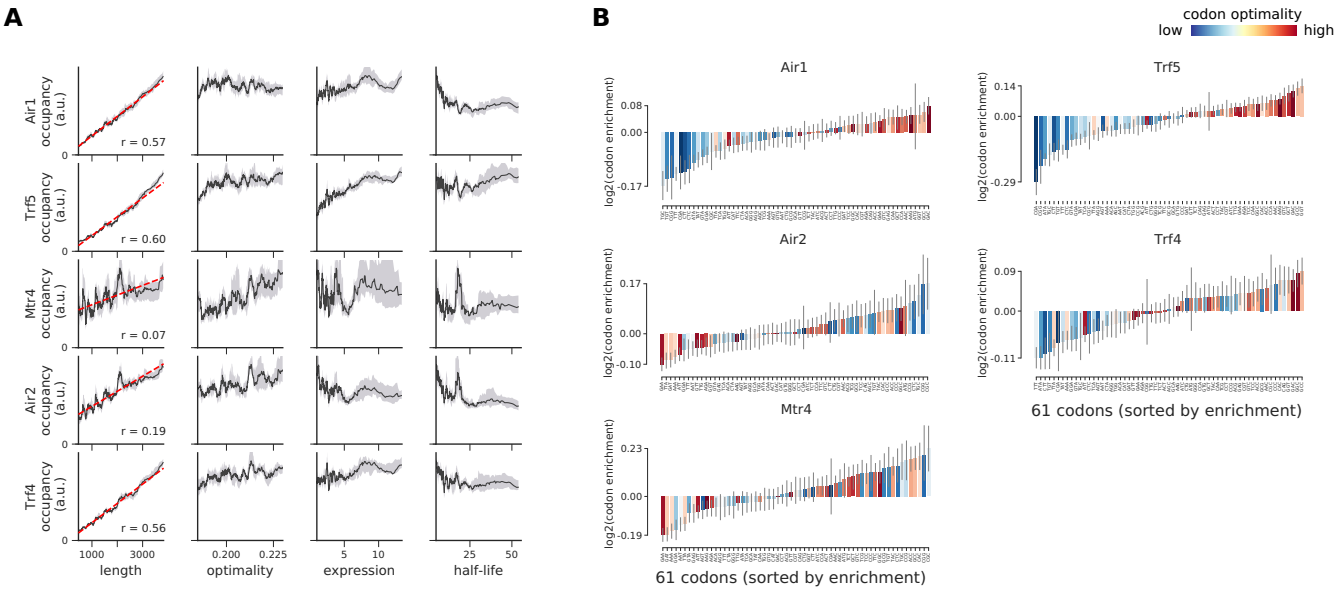


Figure 6—figure supplement 5. Occupancies for components of the TRAMP complex (Air1, Trf5, Mtr4, Air2, and Trf4) compared to transcript length, optimality, expression level, and half-life. (**A**) To understand binding specificity of TRAMP components, the total occupancy of each factor on a transcript is plotted against various transcript features (Gray shading: 95% confidence intervals generated by bootstrapping transcripts). (**B**) Same analysis as in ***Figure 6B***: Codon enrichment shows deviations in codon frequencies of transcripts bound by a degradation factor compared to each codon's frequency on all coding sequences. Each bar is colored according to its codon-optimality with highly optimal codons in dark red and highly non-optimal codons in dark blue. (Gray lines: 90% confidence intervals generated by bootstrapping coding sequences).

DOI: https://doi.org/10.7554/eLife.47040.025

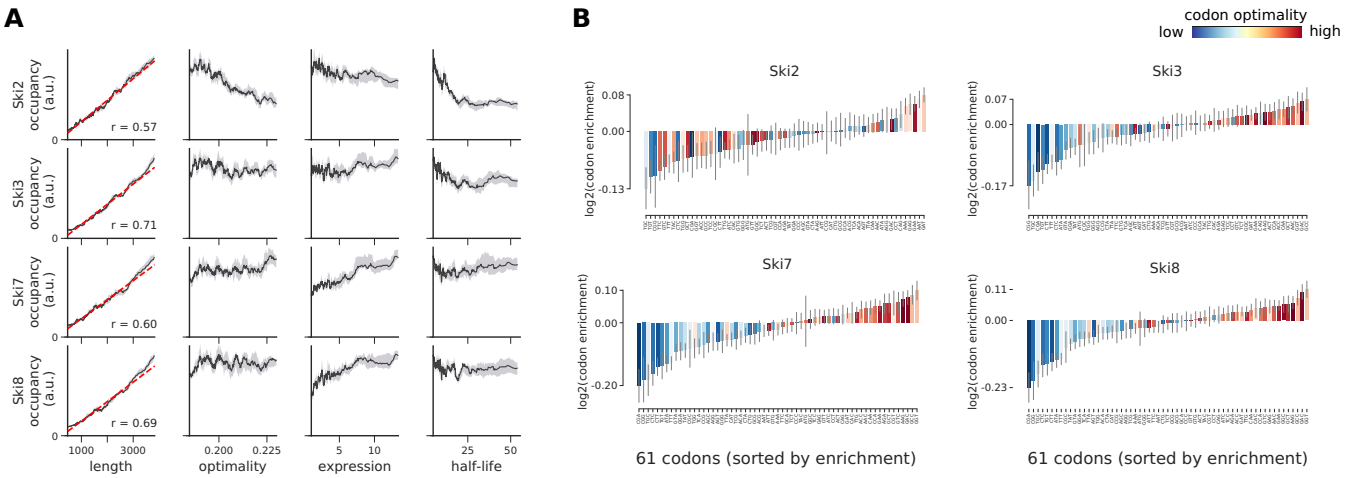


Figure 6—figure supplement 6. Occupancies for components of the Ski complex (Ski2, Ski3, Ski7, and Ski8) compared to transcript length, optimality, expression level, and half-life. (**A**) To understand binding specificity of factors in the Ski complex, the total occupancy of each factor on a transcript is plotted against various transcript features (Gray shading: 95% confidence intervals generated by bootstrapping transcripts). (**B**) Same analysis as in ***Figure 6B***: Codon enrichment shows deviations in codon frequencies of transcripts bound by a degradation factor compared to each codon's frequency on all coding sequences. Each bar is colored according to its codon-optimality with highly optimal codons in dark red and highly non-optimal codons in dark blue. (Gray lines: 90% confidence intervals generated by bootstrapping coding sequences).

DOI: https://doi.org/10.7554/eLife.47040.026

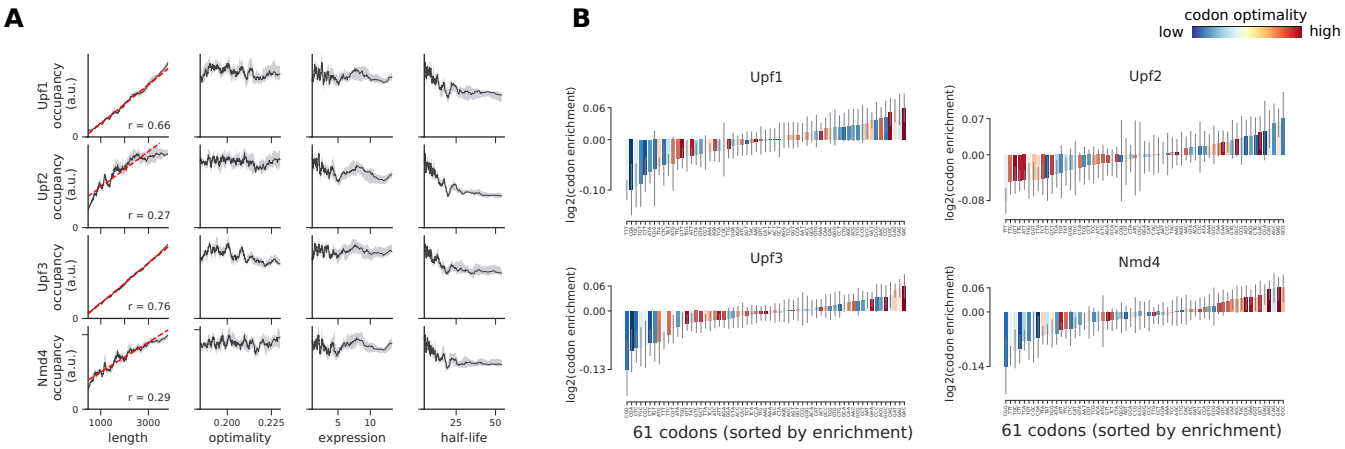


Figure 6—figure supplement 7. Occupancies for components of the NMD pathway (Upf1, Upf2, Upf3, and Nmd4) compared to transcript length, optimality, expression level, and half-life. (**A**) To understand binding specificity of factors in the NMD pathway, the total occupancy of each factor on a transcript is plotted against various transcript features (Gray shading: 95% confidence intervals generated by bootstrapping transcripts). (**B**) Same analysis as in ***Figure 6B***: Codon enrichment shows deviations in codon frequencies of transcripts bound by a degradation factor compared to each codon's frequency on all coding sequences. Each bar is colored according to its codon-optimality with highly optimal codons in dark red and highly non-optimal codons in dark blue. (Gray lines: 90% confidence intervals generated by bootstrapping coding sequences).

DOI: https://doi.org/10.7554/eLife.47040.027

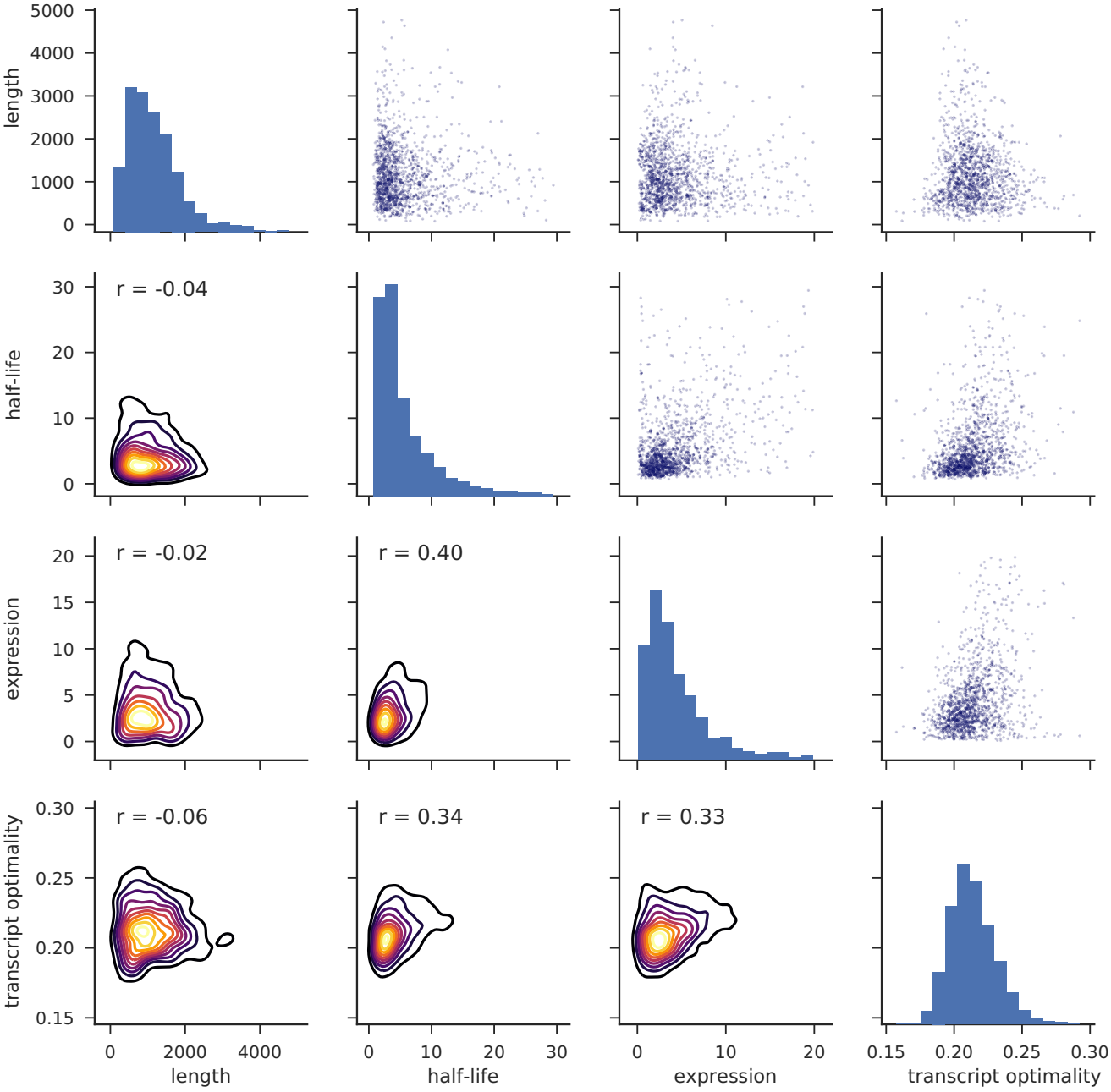


Figure 6—figure supplement 8. Distributions of transcript length, half-life, expression level and transcript optimality for yeast mRNAs. Histograms on the diagonal show distributions of length, half-life (Materials and methods), expression level (***Baejen et al., 2017***) and transcript optimality (***Pechmann and Frydman, 2013***). Pairwise comparisons of features are shown as scatter plots (top right) and kernel density estimates (KDEs) of bivariate densities are shown in the bottom with Pearson correlation values (r) (Materials and methods).

DOI: https://doi.org/10.7554/eLife.47040.028

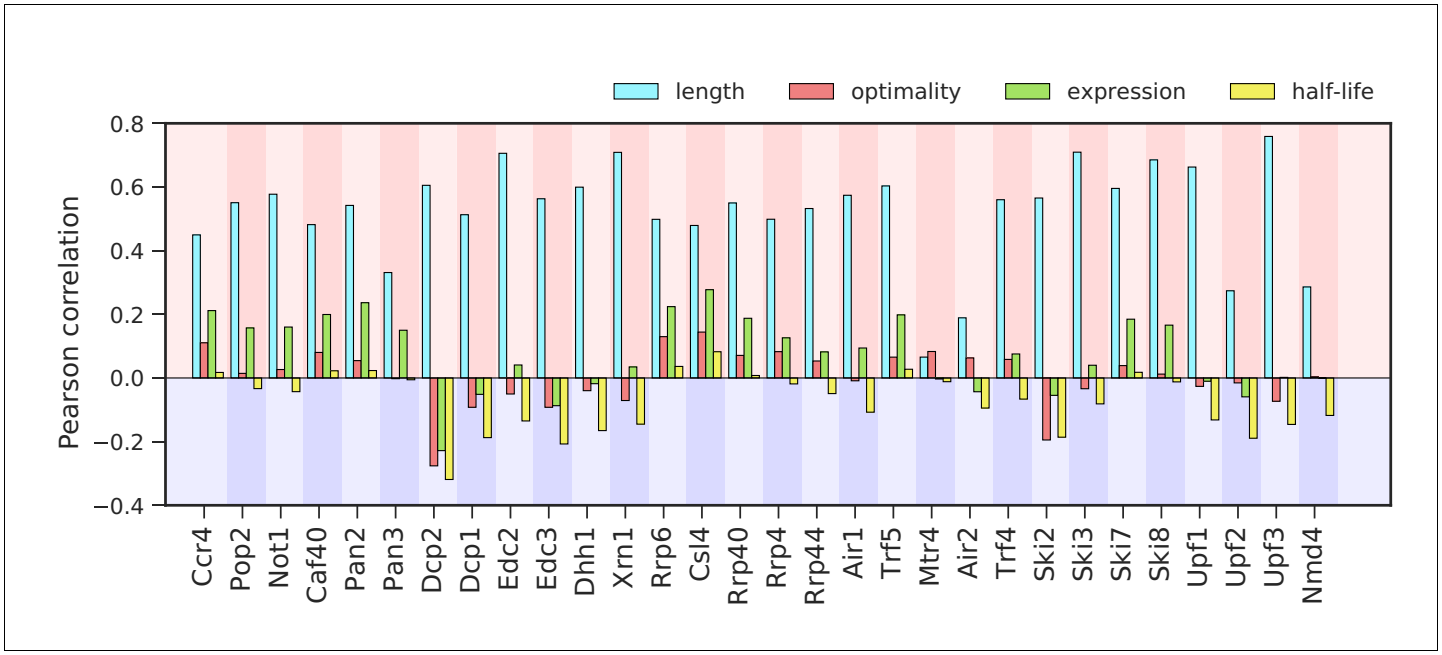


Figure 6—figure supplement 9. Correlation between binding to degradation factors and transcript length, codon-optimality, expression, and half-life. Pearson correlation values between the binding strength of degradation factors (total occupancy over each transcript) and transcript length, transcript optimality (***Pechmann and Frydman, 2013***), expression level (***Baejen et al., 2017***), and half-life derived by multivariate linear regression analysis (Materials and methods).

DOI: https://doi.org/10.7554/eLife.47040.029

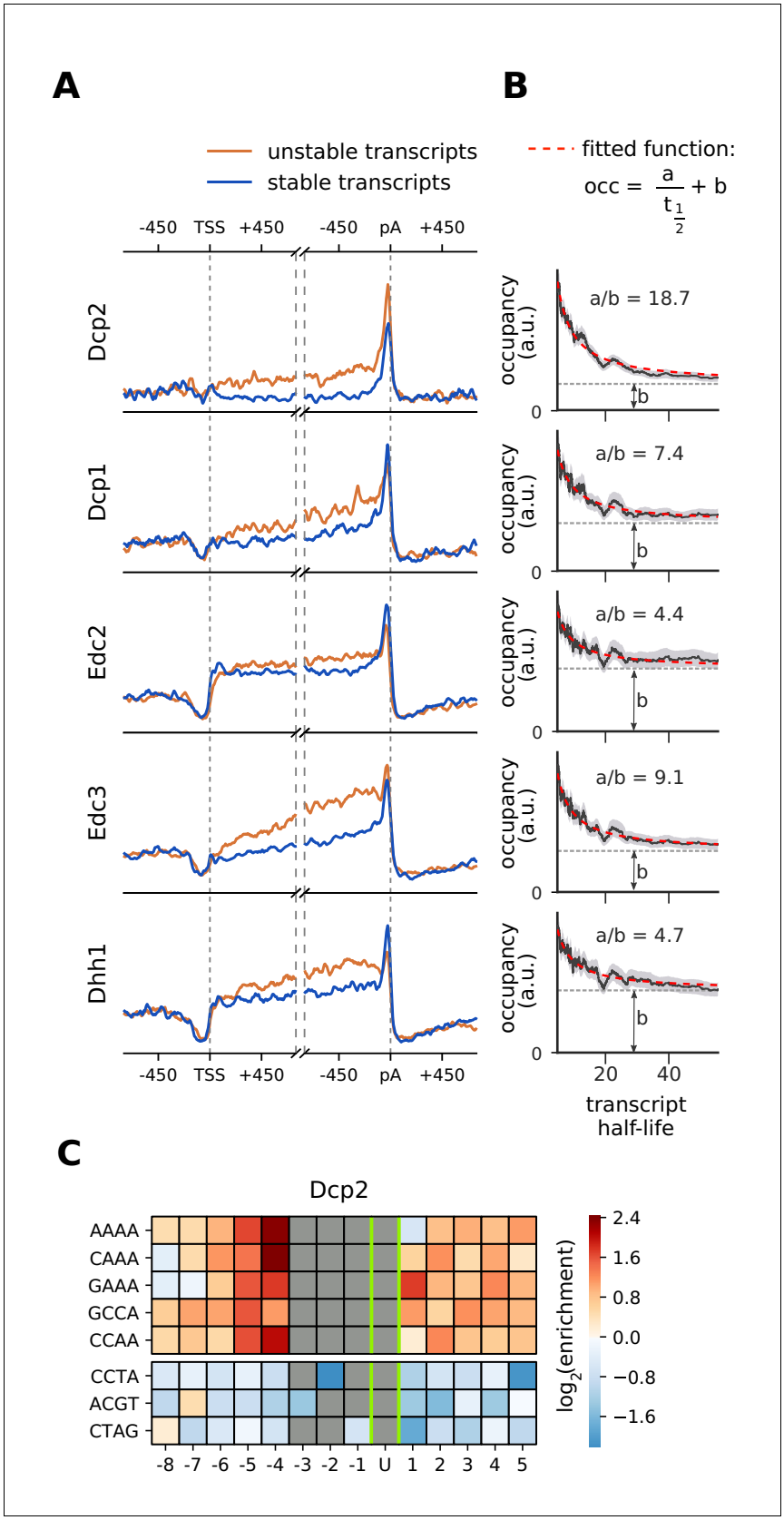


Figure 7. Location and recruitment of the decapping complex Dcp1/Dcp2 and decapping enhancers Edc3, Dhh1, and Edc2. (**A**) Smoothed, transcript-averaged PAR-CLIP occupancy profiles aligned at TSS and pA sites [±750 nt] of

Figure 7 continued on next page

Figure 7 continued

unstable and stable transcripts (first and fourth quantile of half-life distribution, respectively). (**B**) Dependence of total occupancy of factors on the transcripts half-life. The fitting function is plotted in red and the fitted value for b is marked with a dashed gray line. (Gray shade: 95% confidence intervals generated by bootstrapping transcripts). (**C**) Sequence binding preference for the catalytically active subunit of decapping complex (Dcp2), illustrated with the five most enriched and the 3 most depleted 4-mers. The color code shows the log2 enrichment factor of 4-mers around PAR-CLIP cross-link sites [±8 nt]. Dark red represents strong enrichment and dark blue shows strong depletion of a 4-mer. Infeasible combinations are shown with gray. The most highly enriched field is binding AAAAU with the cross-link at the U, which is enriched over random expectation approximately $2^{2.3}$ = 5-fold.
DOI: https://doi.org/10.7554/eLife.47040.030

3 Thermodynamic modeling reveals widespread multivalent binding by RNA-binding proteins.

Publication:

"Thermodynamic modeling reveals widespread multivalent binding by RNA-binding proteins."

S. Sohrabi-Jahromi and J. Söding[†]

(†) corresponding author

bioRxiv (2021), currently under review.

3.1 Author contributions

J. Söding (JS) and **S. Sohrabi-Jahromi (SSJ)** conceptualized the ideas. **SSJ** implemented the software (BMF), created the web server, and performed all the analysis in the manuscript. JS supervised research. **SSJ** and JS wrote the manuscript.

3.2 Code and software availability

BMF source code, documentation, and motif models can be found at `https://github.com/soedinglab/bipartite_motif_finder`. BMF webserver is accessible at `https://bmf.soedinglab.org/`. Web server code can be found at `https://github.com/soedinglab/bmf-webserver`.

BMF user guide explaining the installation process, its main commands, input parameters, and an example pipeline can be found in appendix (section A1).

Thermodynamic modeling reveals widespread multivalent binding by RNA-binding proteins

Salma Sohrabi-Jahromi,[1] and Johannes Söding[1,2,*]

[1]Quantitative and Computational Biology, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany.
[2]Campus-Institut Data Science (CIDAS), Göttingen, Germany.
[*] Correspondence: soeding@mpibpc.mpg.de

January 30, 2021

Abstract

Motivation: Understanding how proteins recognize their RNA targets is essential to elucidate regulatory processes in the cell. Many RNA-binding proteins (RBPs) form complexes or have multiple domains that allow them to bind to RNA in a multivalent, cooperative manner. They can thereby achieve higher specificity and affinity than proteins with a single RNA-binding domain. However, current approaches to de-novo discovery of RNA binding motifs do not take multivalent binding into account.
Results: We present Bipartite Motif Finder (BMF), which is based on a thermodynamic model of RBPs with two cooperatively binding RNA-binding domains. We show that bivalent binding is a common strategy among RBPs, yielding higher affinity and sequence specificity. We furthermore illustrate that the spatial geometry between the binding sites can be learned from bound RNA sequences. These discovered bipartite motifs are consistent with previously known motifs and binding behaviors. Our results demonstrate the importance of multivalent binding for RNA-binding proteins and highlight the value of bipartite motif models in representing the multivalency of protein-RNA interactions.
Availability: BMF source code is available at https://github.com/soedinglab/bipartite_motif_finder under a GPL license. The BMF web server is accessible at https://bmf.soedinglab.org.

Introduction

RNA molecules in the cell are rarely naked but rather covered with numerous RNA-binding proteins (RBPs) (35). These RBPs play a crucial role in regulating the various steps of RNA biochemistry, from RNA maturation and transport, to cellular localization, translation, and degradation (7). RNA molecules can in turn regulate RBP function by altering their stability, interaction partners, and localization (12). These processes require specific binding of RBPs to their target RNAs. RBPs mostly achieve this specificity through RNA-binding domains (RBDs) that engage with specific RNA sequences and\or structures (20). Unraveling the target preferences of RBPs is therefore key to understanding cellular regulation.

Many experimental techniques have emerged to generate systematic maps of protein-RNA interactions. To find *in-vivo* binding sites, many variants of RNA immunoprecipitation (RIP-seq) (9) and cross-linking immunoprecipitation (CLIP-seq), such as PAR-CLIP (10), iCLIP (18) and eCLIP (39), have been proposed. In both approaches, RNAs bound to the immunoprecipitated protein of interest are sequenced and mapped to the genome. Deriving accurate models of binding affinities from *in-vivo* data is problematic because RBP-RNA interactions are influenced by cooperativity and competition with other RBPs, and are additionally influenced by RNA localization, expression, and folding (2). Therefore, techniques have been developed to measure binding affinities *in-vitro*, in isolation from other RBPs, using random libraries of RNA substrates: RNA Bind-n-Seq (RBNS) (19), RNA-compete (30, 5), and high-throughput RNA-SELEX (HTR-SELEX) (13).

A wide range of motif discovery tools have been developed to learn models of sequence- and secondary structure-dependent binding affinties of RBPs based on datasets of sequences bound in-vitro or in-vivo by an RBP of interest (15, 23, 38, 25). More recently, a new wave of algorithms have been introduced that use deep neural networks to predict RBP binding sites (1, 3, 28, 8). While these deep learning methods have promising accuracy in predicting RBP binding, interpreting what these networks have learned remains challenging. Moreover, with the increasing number of model parameters and network complexity, the risk grows that such models could also learn experimental biases in the datasets. This is particularly problematic for RBPs, since many of them show short and degenerate sequence preferences. Moreover, RBPs often bind low-complexity untranslated regions in the RNA (6), unlike transcription factors, which usually bind to more complex sequence motifs and have higher binding specificities. RBPs have further been shown by spaced k-mer counting approaches to

often bind with multiple RNA-binding domains two separated cores with usually similar or identical motifs (6, 13). A recent deep learning software is the only available one capable of learning distance dependent motif pairs (29).

In this work, we present Bipartite Motif Finder (BMF), a tool for learning bipartite RNA motifs in RNA-ptotein interaction datasets. To accurately model the binding affinity of RBPs possessing domains with similar core motifs of low-information content that bind to RNAs with a relatively high density of the core motifs, BMF sums up the contribution of all alternative binding conformations, and not just the best binding configuration. To the best of our knowledge, BMF is the first approach that adopts a thermodynamic viewpoint to RBP *de novo* binding motif discovery. We demonstrate that BMF is able to detect short and degenerate motifs and to learn the spatial relationship between them. We furthermore show that around half of RBPs manifest multivalent binding with a preferential linker distance between the two binding sites.

Benchmarking the performance of learned binding site models by cross-validation can be problematic when testing methods that train highly parameterized models such as deep neural networks, as these methods can learn biologically irrelevant sequence biases inherent to the experimental method. To compare BMF to existing tools and assess their capacity for learning relevant motif sequences that predict binding events in the cell, we built a cross-platform validation benchmark, training models on HTR-SELEX data and testing on *in-vivo* CLIP data. Despite the many complicating effects *in vivo*, we find that the motif and distance preferences learned by BMF can predict RBP binding in the cellular context and that high-quality motifs learned *in vitro* are often very similar to the motifs learned on *in-vivo* data . Moreover, BMF can predict binding sites on par with or even better than existing tools.

Methods

Most RBPs can bind RNA using several structured RBDs and often also using disordered regions, some of which contain typical RGG/RG and RS motifs, which can modulate RNA-binding activity (21, 4, 27). Furthermore, many RNA-binding proteins dimerize or homo- and hetero-oligormerize. This effectively leads to two and more RNA-binding domains binding cooperatively to RNA molecules. Here, we present Bipartite Motif Finder (BMF), a motif search tool and algorithm to describe the sequence specificity of monovalently and multivalently binding proteins or protein complexes.

Thermodynamic model for bivalent RNA binding

We consider the simple case in which the RBP consists of two RBDs, A and B (Figure 1A). We describe the binding of proteins at concentration c_{AB} to a single, specific RNA sequence $\mathbf{x} = (x_0 \dots x_{L-1}) = x_{0:L-1}$ composed of nucleotides x_i. We consider not only the most likely binding configuration but rather all possible binding configurations, involving zero, one or more proteins bound to the RNA (Figure 1B). According to Boltzmann's law, each binding configuration $\mathbf{c}$ has a probability $p(\mathbf{c})$ proportional to its so-called *statistical weight* $e^{(-E(\mathbf{c})-T\Delta S(\mathbf{c}))/k_{\mathrm{B}}T}$, where $F(\mathbf{c}) = -E(\mathbf{c}) - T\Delta S(\mathbf{c})$ is the free energy composed of the binding enthalpy $-E(\mathbf{c})$ and a part related to the change in entropy $\Delta S(\mathbf{c})$ between the completely unbound and bound states. To obtain probabilities, the statistical weights need to be normalized at the end by dividing by their total sum, the partition sum $Z(\mathbf{x})$.

The change in entropy due to the binding of a single protein that is present at concentration c_{AB} is equal to its chemical potential, which is $\Delta S = k_{\mathrm{B}} \log c_{\mathrm{AB}}$. In the following, we compute all energies in units of $k_{\mathrm{B}}T$, so we set $k_{\mathrm{B}}T = 1$. In our model, the concentration $c_{\mathrm{B}}(d)$ of the downstream domain B at the RNA depends on the distance d to the binding site of the upstream domain A (see next subsection).

We compute the statistical weights of all binding configurations iteratively using dynamic programming. We split the configurations into two sets, A and B, and define $Z_{\mathrm{A}}(i)$ to be the sum of statistical weights of all binding configurations on the RNA up to position i, $x_{0:i}$, for which domain A is bound at position $i-k+1$ to i, where k is the length of RNA bound by the domains. Similarly, we define $Z_{\mathrm{B}}(i)$ to be the sum of statistical weights of all binding configurations on the RNA sequence $x_{0:i}$ for which no domain is bound or domain B is bound with its right edge upstream of or at position i. With the knowledge of $Z_{\mathrm{A}}(i')$ and $Z_{\mathrm{B}}(i')$ for $0 \leq i' < i$, we can compute $Z_{\mathrm{A}}(i)$ and $Z_{\mathrm{B}}(i)$ (Figure 1B):

$$Z_{\mathrm{A}}(i) = \left(Z_{\mathrm{B}}(i-l) \; + \sum_{j=0}^{i-l} Z_{\mathrm{A}}(j) \right) c_{\mathrm{AB}} e^{-E_A(x_{i-k+1:i})}, \quad (1)$$

$$Z_{\mathrm{B}}(i) = Z_{\mathrm{B}}(i-1) + \sum_{j=0}^{i-l} Z_{\mathrm{A}}(j)\; c_{\mathrm{B}}(i-k-j)\; e^{-E_B(x_{i-k+1:i}))} + Z_{\mathrm{B}}(i-l)\, c_{\mathrm{AB}}\, e^{-E_B(x_{i-k+1:i})}\,, \quad (2)$$

where $E_A(x_{i-k+1:i})$ and $E_B(x_{i-k+1:i})$ represent the binding energies of domains A and B to the RNA sequence $x_{i-k+1:i}$. The concentration of the single B domain, defined as expected number of B per volume, is simply its probability density. The dynamic programming is initialized using

$$Z_{\mathrm{A}}(i) = 0 \text{ for all } i \in \{0, ..., k-2\}, \quad (3)$$

$$Z_{\mathrm{B}}(i) = 1 \text{ for all } i \in \{0, ..., k-2\}\,. \quad (4)$$

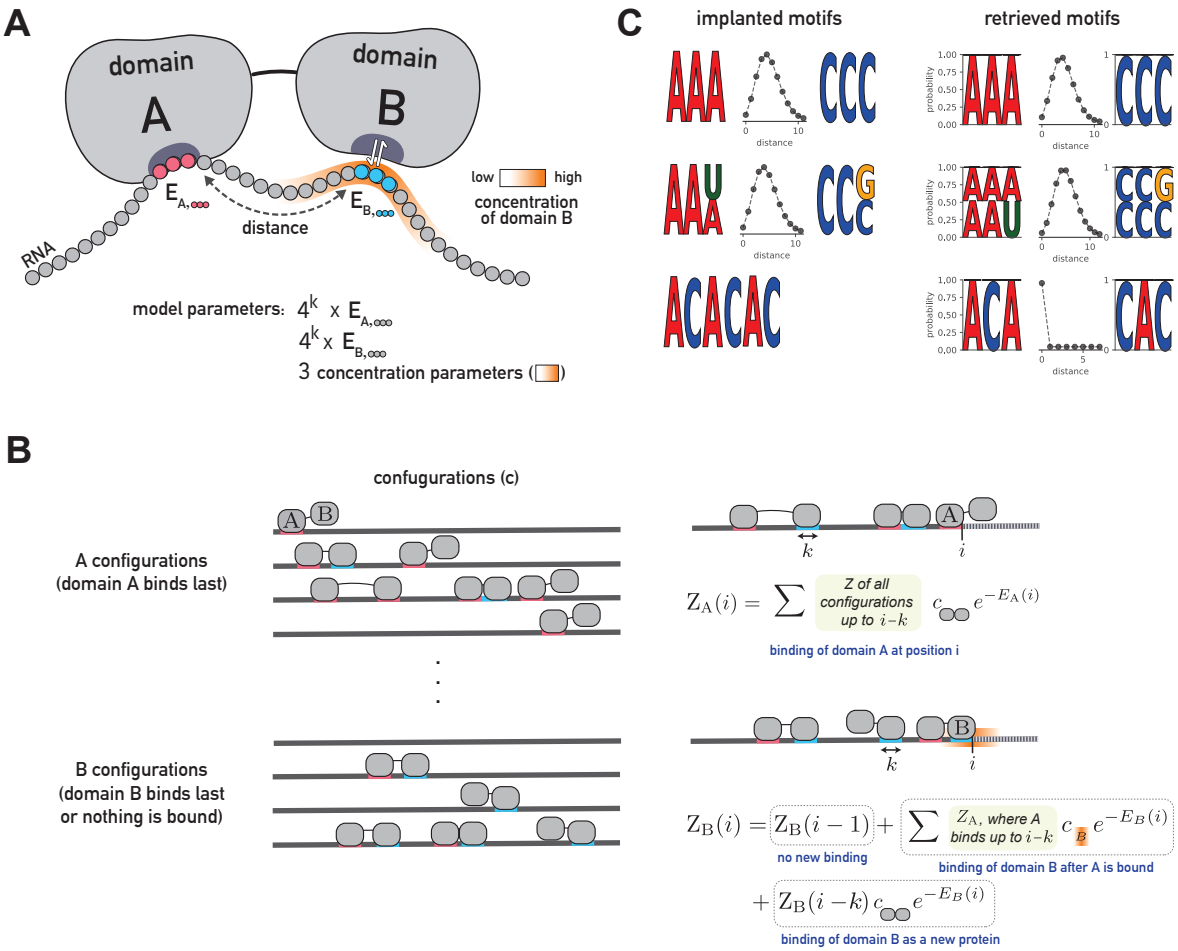


Figure 1: **BMF can learn multivalent binding preferences for RBPs. (A)** RBP-RNA interaction model for a protein with two RNA-binding domains. BMF optimizes the binding energies of each domain to all possible RNA k-mers ($k = 3$ here) and learns the distance distribution between the motif cores. BMF models the high RNA local concentration at the second binding site, when the first domain is bound to the RNA. **(B)** BMF calculates binding probabilities for all binding configurations of one or several proteins to the RNA sequence. $Z_{\mathrm{A}}(i)$ is the sum of statistical weights of all binding configurations on the RNA up to position i, for which domain A is bound at position i. Similarly, $Z_{\mathrm{B}}(i)$ is the sum of statistical weights of all binding configurations on the RNA subsequence for which no domain is bound or domain B is bound with its right edge upstream of or at position i. Z_{A} and Z_{B} are calculated iteratively (right panel). The first term in the second equation accounts for configurations for which position i is not bound by anything, the second term accounts for configurations for which domain A of the same protein is bound at j (as seen in the example illustration) and the last term accounts for configurations for which domain B binds whose A domain is not bound upstream of i. **(C)** BMF recovers the correct RNA motifs implanted in synthetic datasets for all tested cases. Here and in the following figures, the two learned core motifs are visualized by plotting the energies of the top five k-mers, converted to k-mer probabilities according to Boltzmann's law and normalized to 1.

The first equation follows from requiring all k positions in the binding motif to be part of sequence $x_{0:L-1}$. The second equation follows from the fact that $Z_{\mathrm{B}}(i)$ for $i < k-1$ sums up only the statistical weight of the unbound configuration.

The partition sum $Z(\mathbf{x})$ for RNA sequence $\mathbf{x}$ is the sum of statistical weights of all configurations,

$$Z(\mathbf{x}) = Z_{\mathrm{B}}(L-1) + \sum_{i=0}^{L-1} Z_{\mathrm{A}}(i)\ . \tag{5}$$

The probability for an RNA to not be bound by any protein (neither A nor B domains) is just the statistical weight of the unbound configuration, set to 1, times the normalization factor $1/Z(\mathbf{x})$, so the probability for a RNA $\mathbf{x}$ to be bound by a protein is $p(\mathrm{bound}|\mathbf{x}) = 1 - 1/Z(\mathbf{x})$.

By taking the partial derivatives of equations (1) and (2) with respect to the model parameters (Supplementary Methods), we obtain update equations for the partial derivatives with which we can in turn compute the partial derivatives of $Z(\mathbf{x})$, $p(\mathrm{bound}|\mathbf{x})$, and the log likelihood in equation (8). These allow us to find optimum model parameters by gradient-based maximization of the log-likelihood.

Motif model of a single RNA-binding region

Position weight matrices (PWMs) and Bayesian Markov models (BaMMs) have been used to represent RBP binding preferences through positional or conditional proba-

bilities of observing each nucleotide at a given position (11, 34). Since RBPs are known to bind shorter and more repetitive sequences, we learn binding energies for all 4^k k-mers at each motif core, $E_A(k\text{-mer})$ and $E_B(k\text{-mer})$. The length k of the motif can be set by the user.

Model for the effective concentration $c_B(d)$

Spaced k-mer analyses on high-throughput RNA-binding datasets pointed to a length preference of the RNA linker connecting two motif cores (13, 6, 32). The concentration of domain B after domain A binds the RNA molecule is equal to its probability distribution. While according to the flexible chain model of the RNA fragment the concentration should be a Gaussian distribution centered on domain A (31), for short RNA linkers the concentration can peak some distance away from domain A. To describe multivalent binding for both short-range and long-range co-occurrence of motif sequences, we model the effective concentration at the second binding site with a negative binomial (NB) distribution,

$$c_B(d) = c_{AB} + S \cdot \binom{d+r-1}{d} \cdot p^r (1-p)^d, \qquad (6)$$

where d represents the the number of nucleotides between the binding sites of A and B on the RNA, and r and p are parameters of the negative binomial distribution. The total concentration of B is the cellular concentration (c_{AB}) plus $c_B(d)$, the local concentration of B linked to a bound A. We scale the negative binomial with the factor S as a conversion to protein concentration values. Since only the ratio between S and c_{AB} determine the binding dynamics, we fix c_{AB} to one and optimize our bipartite model for S, r, and p.

Parameter initialization

The absolute values of the energy parameters in our model do not reflect the physical binding energies, however their relative values determine the probability of binding to a given sequence. We therefore draw initial energy parameters randomly (in units of k_BT) from a normal distribution with the average of 12 and standard deviation of one. The initial value of 12 k_BT was chosen based on experimentally determined binding energies (43) and additionally ensures that the algorithm does not overflow. The scaling factor S is initialized as 10^4, The spacer parameter r is drawn from a uniform distribution from one to five and p is randomly drawn between zero and 0.5.

Likelihood estimation for HTR-SELEX measurements

In HTR-SELEX experiments (and similarly for bind-n-Seq), we have input (background) library sequences $\mathbf{x} \in \mathcal{X}^{bg}$ and sequences enriched after competitive binding, $\mathbf{x} \in \mathcal{X}^{+}$. We denote with $p_b(\mathbf{x})$ the fraction of sequence $\mathbf{x}$ in the input library. To find a sequence in $\mathbf{x} \in \mathcal{X}^{+}$, it must have first been present in the input library (probability $p_b(\mathbf{x})$) and then have been bound to the RNA (probability $p(\text{bound}|\mathbf{x})$). The probability to find a sequence $\mathbf{x} \in \mathcal{X}^{+}$ after the selection is therefore, according to Bayes' theorem,

$$p(\mathbf{x}|\text{bound}) = \frac{p(\text{bound}|\mathbf{x})\, p_{bg}(\mathbf{x})}{\sum_{\mathbf{x}' \in \mathcal{X}^{bg}} p(\text{bound}|\mathbf{x}')\, p_{bg}(\mathbf{x}')}, \qquad (7)$$

and, using $p(\text{bound}|\mathbf{x}) = 1 - 1/Z(\mathbf{x})$, the log-likelihood is

$$\begin{aligned} LL &= \ln \prod_{x \in \mathcal{X}^{+}} p(\mathbf{x}|\text{ bound}) \\ &= \sum_{\mathbf{x} \in \mathcal{X}^{+}} \left(\ln p_{bg}(\mathbf{x}) + \ln\left(1 - \frac{1}{Z(\mathbf{x})}\right)\right) \\ &\quad - N^{+} \ln \sum_{\mathbf{x}' \in \mathcal{X}^{bg}} p_{bg}(\mathbf{x}')\left(1 - \frac{1}{Z(\mathbf{x})}\right). \end{aligned} \qquad (8)$$

Parameter optimization

We learn the model parameters by maximizing the likelihood function (eq. 8). For an efficient optimization using stochastic gradient descent, we computed the partial derivative of the likelihood function with respect to all of the model parameters (Supplementary Methods). For parameter optimization, we used ADAM (16) with hyperparameters $\alpha = 0.01$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-8}$, and a minibatches size of 512. We parameterized $r = e^{\rho}$ and $p = 1/(1 + e^{-\pi})$ to ensure that r and p stay within bounds. Optimization was terminated when 1000 iterations were reached or when the variation v_θ for the best bound k-mer of each domain as well as for p and r fall under a threshold of 0.03. The variation for the parameter θ up to iteration t was defined as $v_\theta = (\max\{\theta_{t\text{-}4:t}\} - \min\{\theta_{t\text{-}4:t}\})/\theta_t$.

Evaluating the performance of BMF on synthetic data

In order to evaluate BMF's ability to learn bipartite motifs, we generated two sets of 2000 RNA sequences, an artificial input set and an enriched set. For the enriched set, we inserted the first core of the simulated bipartite motif at random positions. The second core was inserted with a linker length drawn from a binomial distribution with a specific p and r. We ran BMF 10 times with random parameter initializations to assess its robustness.

HTR-SELEX datasets

We obtained 177 HTR-SELEX datasets of 86 distinct factors from (13). We used sequences of the input library

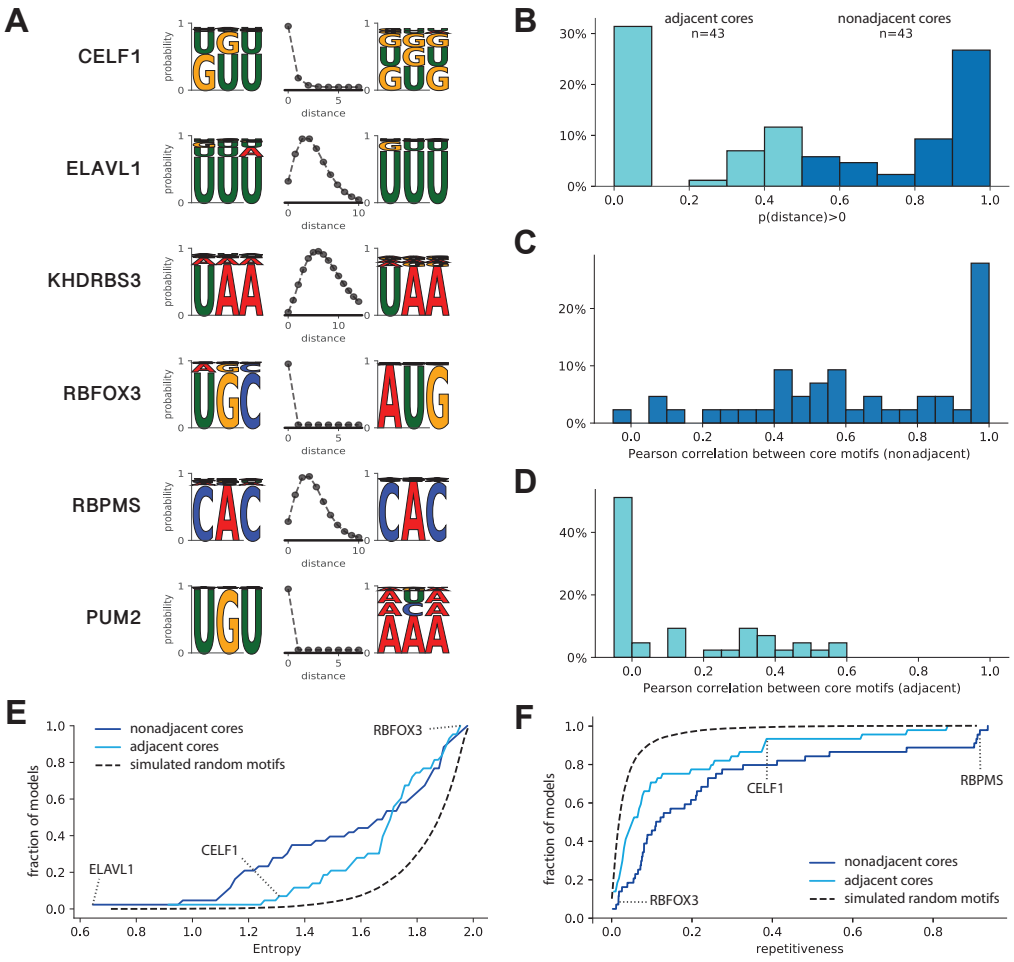


Figure 2: **Many RBPs are multivalent, bind low-complexity sequences and often bind two similar motif cores. (A)** Examples of motifs that represent a wide range of binding modes, learned by BMF on HTR-SELEX data. When the RBP has a larger motif than allowed by the core size (3 here), the distance between cores is learned to be zero to accommodate a longer binding sequence (e.g. CELF1, RBFOX3, and PUM2). **(B)** Distribution of the probability of the spacer length between the two motif cores to be above 0 . As seen in the examples in A, most RBPs either clearly bind adjacent cores (distance= 0, turquoise) or have a multivalent binding mode with two nonadjacent cores (dark blue). **(C)** and **(D)** Similarities between binding preferences of the two cores for RBPs with adjacent cores (turquoise) or multivalent nonadjacent cores (dark blue), according to panel B. **(E)** Cumulative distribution of the entropy of BMF models for all RBPs in the HTR-SELEX dataset. In general the optimized bipartite motif models have much lower complexity than randomly generated bipartite models (dashed black line). **(F)** Cumulative distribution of *"sequence repetitiveness"* of BMF models for all RBPs in the HTR-SELEX dataset. Overall, BMF models are more often repetitive that those of randomly generated bipartite models (dashed black line).

and the last cycle to train BMF. Even though our model describes one cycle of selection, the retrieved motifs were more prominent in the later cycles. Moreover, the cross-platform validation discussed below resulted in slightly better performance for all the tools when choosing the input and last cycles for motif detection in comparison to second and third cycles. Whenever several experimental or technical replicates were available, we built a separate model for each replicate and averaged the corresponding metric over all replicates of an RBP at the end. We used BMF's default hyper-parameters throughout the manuscript.

Cross-platform validation of in-vitro motifs

Each experimental technique for measuring RNA binding has its own biases. When measuring the quality of predictions of motif models by cross-validation, methods can learn these biases to distinguish bound from background sequences. Highly parameterized models could learn such subtle, complex biases. These platform-dependent biases can be a result of library preparation, amplification, or can depend on the type and concentration of RNase that is used (17, 26). There have been efforts to reduce the effect of such biases when training motif models, e.g. by learning binding models for many RBPs at the same time (8). In order to ensure that BMF does not over-train on the *in-vitro* HTR-SELEX data, we performed cross-platform validation: We trained BMF on HTR-SELEX datasets and used the resulting models to predict binding sites in *in-vivo* CLIP data.

We collected eCLIP datasets of 15 RBPs (40) and PAR-CLIP datasets of 10 RBPs (24) for which we also have HTR-SELEX data. We used the pre-processed CLIP

peaks as enriched sequences. Since the PAR-CLIP dataset contained larger numbers of peaks, we restricted our analysis to the top 2000 reported binding sites per RBP. For each eCLIP and PAR-CLIP dataset, we created a background set of the same size by drawing random PAR-CLIP or eCLIP peaks of other factors measured with the same technique. We applied a sliding window with length of 50 and a stride of 20 to generate same-size fragments that fully cover each peak. The prediction scores were averaged over these fragments when the region was longer than 50 bases. We compared our simple model with deep learning approaches, the popular RBP binding predictors iDeepE (28) and GraphProt (23). iDeepE uses deep learning to predict RBP binding, while GraphProt's model is based on Support Vector Machines (SVMs).

BMF software and web server

The BMF command-line tool offers three commands: (1) Learning a BMF model given enriched and background sequences. Output is a BMF model file. (2) Bipartite motif visualization, given the BMF file learned in step 1. (3) Predicting binding scores for new sequences with the BMF model trained in the first step. The first two functionalities (*de novo* motif discovery) are also available on the BMF web server.

Results

We present BMF, a method for de-novo discovery of RNA-binding motifs that uses a bipartite motif model capable of learning multivalent binding specificities among RBPs. BMF models the protein binding with up to two domains to its RNA substrate. We assume that due to the structure of the RBP (or RBP complex), the distance between the two binding sites is spatially constrained. BMF therefore consists of two short sequence motif models and a distance probability distribution (Figure 1A). Binding with just one domain is modeled using a distance distribution peaked at 0 base pairs. In the following sections we demonstrate that this model can reliably detect bipartite motifs in synthetic and real sequences, and we evaluate its performance at identifying binding sites compared to other models of RBP binding in HTR-SELEX, PAR-CLIP and eCLIP datasets.

BMF accurately discovers implanted synthetic motifs

To test BMF's ability to learn bipartite motifs, we generated 2000 artificial sequences containing first an AAA and then a CCC with a distance distribution of around 3 to 5 bases between them (Figure 1B, top). BMF retrieved the implanted motifs and spacer distribution accurately. The results were similarly accurate when sequence degeneracy was introduced by flipping the last base (Figure 1B, middle), or when implanting the repeat sequence ACACAC (Figure 1B, bottom). This demonstrates that BMF can not only reveal multivalent specificities but can also recover longer sequence motifs by placing the two cores adjacent to one another.

The log-likelihood increases during stochastic gradient descent and the optimization terminates when the log-likelihood has reached a plateau (Figure S1A). The distance parameters and the binding energies of k-mers in the motif cores all reach a plateau before termination (Figure S1B,C). To test robustness to parameter initialization, we ran BMF ten times with random initial parameter values and verified that the k-mer energies and distance parameters match across all runs. (Figure S1D,E).

Most RBPs show multivalent binding, often to multiple occurrences of the same motif

We applied BMF to 177 HTR-SELEX datasets consisting of 86 distinct RBPs to investigate the importance of multivalent binding in the formation of RBP target specificity. BMF detected bipartite binding for many RBPs including ELAVL1, KHDRBS3, and RBPMS (Figure 2A). Interestingly, BMF restricted the distance of the motif cores strictly to zero when the RBP binds repeat sequences (e.g. CELF1 binding GU repeats) or when the RBP binds a longer RNA sequence that requires a longer motif core (e.g. RBFOX3 binding UGCAUG, and PUM2 binindg UGUANA). The sequence and spacing preferences were also reproducible across experimental replicates (Figure S2), and match for proteins that belong in the same family (Figure S3). All 177 BMF models with core lengths of 3-5 can be found at BMF's GitHub repository. These results show that BMF can identify bipartite motifs in HTR-SELEX data.

We then looked for the frequency of such multivalent, bipartite motifs and calculated the probability of observing the two core motifs at distances beyond zero for each motif model (Figure 2B). At two extremes, this probability would be zero for RBPs like RBFOX3, which consist of a larger binding sequence, and one for RBPs like KHDRBS3, which prefer a larger spacer between the motif cores. Interestingly, the majority of RBPs lie at the two extremes, and about half of them show a bipartite binding behavior. This ratio is higher than estimated in previous studies, which were based on k-mer counting approaches (13, 6). The number of bipartite motifs could be furthermore underestimated as some RBPs show bipartite binding only when BMF's core size is increased to four or five nucleotides (Figure S4). Overall, these results highlight the importance of multivalent binding as a common strategy to achieve high specificity despite having individually small and weak binding sites.

We noted that many motif models (like ELAVL1 and KHDRBS3) have similar sequence preferences on both cores. We quantified their similarity by the Pearson correlation between the probabilities of observing each of the 4^k k-mers. As expected from the individual examples, the core motifs are mostly similar for RBPs that exhibit bipartite binding (Figure 2C) as opposed to adjacent motif cores (Figure 2D). This demonstrates that RBPs have often evolved to bind multiple occurrences of the same or similar short sequence motifs, either using multiple same-chain RNA-binding domains or by homodimerization and oligomerization.

RBPs often bind low-complexity and repetitive sequences

It has been shown that RBPs bind sequences of lower complexity than DNA-binding transcription factors (6, 36). This can be seen at its extreme for some of our binding models, which are composed of only one to two types of nucleotides (Figure 2A). Looking at all 78 RBP binding models, we observed that many proteins bind repetitive sequences or have the same simple k-mer affinities for each of their valencies. In order to quantify this, we calculated the entropy of the motif sequences as a measure of sequence complexity (Figure 2E, Supplementary Methods)(6). For highly complex sequence affinities (e.g. RBFOX3), the entropy gets close to two, while this value is closer to zero for degenerate and repetitive sequences (e.g. ELAVL1). A similar trend is visible when quantifying the repetitiveness of BMF models, resulting in high scores when both cores consist of mono- or di- nucleotide repeats (Figure 2F, Supplementary Methods). Overall, more than half of RBP motifs show levels of degeneracy that are highly unlikely in artificially generated random motif models. This binding preference towards low complexity sequences fits to the previous observation that bipartite motifs tend to bind multiple occurrences of the same sequence.

Including all binding configurations and cooperativity enhances the accuracy of RBP binding predictions

To assess the value of cooperativity and multivalency, we compared BMF to a 6-mer motif model which scores the sequences by finding the best binding site (Figure S5). Interestingly, for all RBPs but particularly for those that show bipartite binding, BMF's performance is superior to that of the 6-mer enrichment model. This highlights the value of two distinct BMF features: considering all binding configurations, and including the cooperative effect of multi-domain binding.

In-vitro bipartite models learned by EMF can predict in-vivo binding

Experimental techniques for measuring RNA binding have individual biases that can be learned by motif discovery tools. This is particularly problematic when evaluating computational methods with many model parameters that can capture complex structures in their input datasets (8). Cross-platform validation, i.e. using binding models trained on an experimental dataset to predicting binding sites in another experimental platform ensures a fair assessment of the quality of motif models. We therefore trained models on HTR-SELEX data to predict binding sites on sequences derived from PAR-CLIP and eCLIP experiments (40, 24). We compared the performance of BMF to iDeepE (28) and Graphprot (23) (Figure 3A-C). iDeepE is a deep learning tool and Graphprot is based on support vector machines. Thanks their more complex architecture and higher number of parameters, both models are able to learn more complex aspects of the training data, while Graphprot additionally takes the RNA structure as an input. Interestingly, despite these advantages, BMF showed a competitive prediction quality as measured by the area under the receiver operating characteristic curve (AUROC), with a better median AUROC than iDeepE and GraphProt. Simlar results are obtained when replacing AUROC with the area under the precision recall curve (AURPC, Figure S6).

Interestingly, generally performance of BMF is best for $k = 3$, although it changes little between core size of $k = 3$, 4 or 5 (Figure 3A, Figure S7). For some RBPs increasing the core size reduced the predictive power for the resulting models. This could be due to over-fitting on biases of the HTR-SELEX data and might be a reason for why the more highly parameterized RNA motif models of GraphProt and iDeepE often do not perform as well as the simpler ones of BMF. On the other hand, longer BMF models, as well as iDeepE and GraphProt, could better learn binding preferences for factors such as CSTF2T that bind more complex RNA sequences. To summarize, BMF can capture RBP specificities with reduced risk of overfitting.

To see whether the core spacing of HTR-SELEX motif models exist in *in-vivo* data, we trained BMF models on the CLIP data and compared them to their *in-vitro* counterparts. Interestingly for the models that were learned well on the HTR-SELEX data (tool-averaged AUROC ≥ 0.8), both the motif core sequences and their distance distribution match between the two experimental platforms (Figure 4). The sequence and\or preferences do vary for other factors with lower AUROC values (Figure S8,S9).

A comparison of the AUROC values from the cross-validated HTR-SELEX data (Figure S5) and those from the cross-platform validation shows a correlation be-

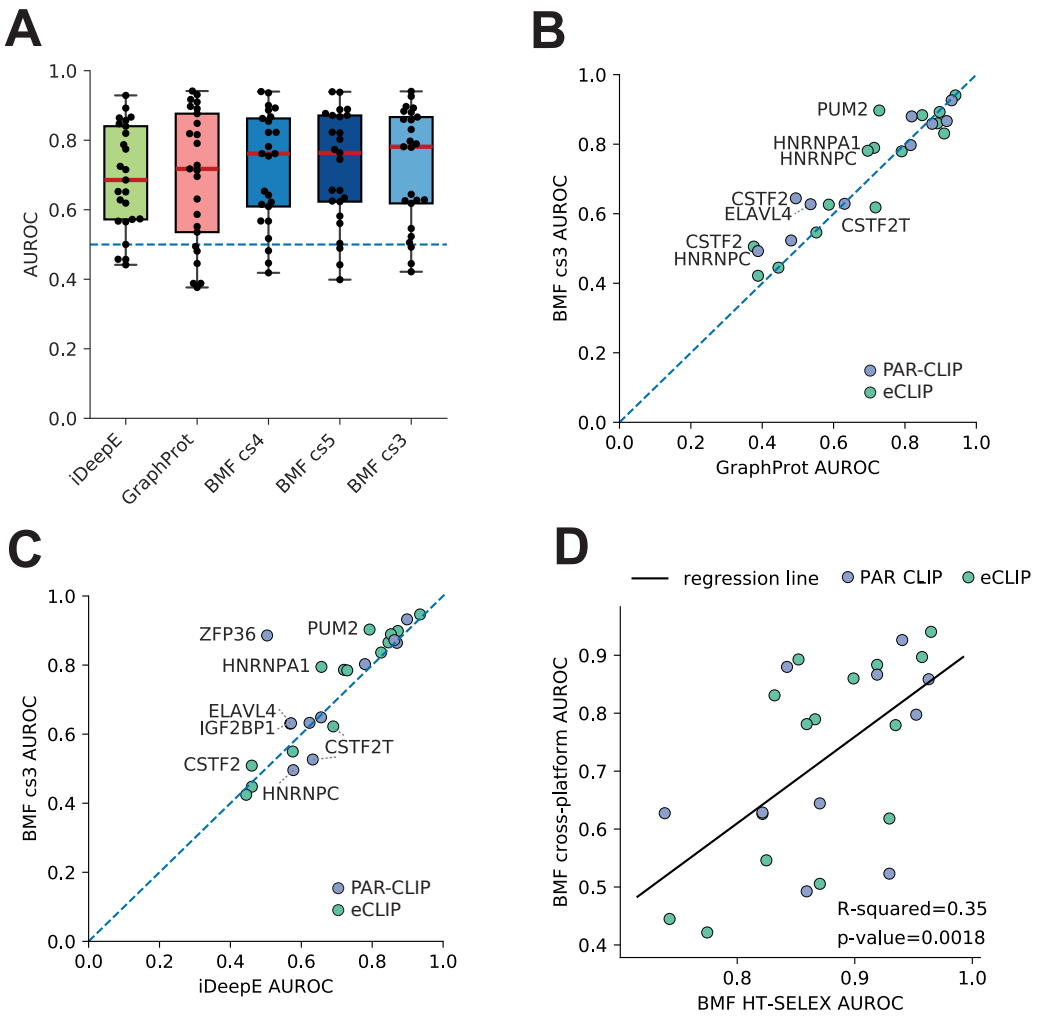


Figure 3: **Cross-platform validation shows *in-vitro* BMF motifs can predict *in-vivo* binding sites in transcriptomes.** We used BMF, iDeepE, and GraphProt to identify eCLIP and PAR-CLIP RBP binding sites after training their motif models on HTR-SELEX datasets. **(A)** AUROC distribution for iDeepE, GraphProt and BMF with motif sizes ranging from 3 to 5. The tools are sorted based on their median AUROC performance. The values for each RBP dataset is shown with a black dot. **(B)** and **(C)** AUROC from BMF (core size 3) compared to iDeepE and GraphProt. **(D)** BMF AUROC values from cross-validated HTR-SELEX analysis can predict cross-platform performance. Both BMF models are built with core size 3. Linear regression line is marked with black. In all plots AUROC values are averaged over all replicate combinations wherever replicates were available.

tween BMF motif quality and its performance in the cross-platform benchmark (p-value=0.0018, Figure 3D). It could help explain why some HTR-SELEX models fail at predicting binding to new sequences, possibly as they have little sequence preference for their target RNA or due to the absence of this information in the HTR-SELEX data. Overall, this shows that BMF can be used to learn RNA motifs from *in-vitro* data to predict binding sites of the protein in the cell despite numerous factors counfounding binding *in vivo*.

Discussion

We present BMF, the first bipartite motif model to describe multivalent binding preferences in RBPs. The motif models learned on *in-vivo* and *in-vitro* datasets imply the following multipartite binding strategy is common – adapted by about half of RBPs in our datasets – to bind their target RNA molecules: First, these RBPs bind multiple short ($3-4$ nt) RNA segments simultaneously and cooperatively with their multiple RBDs, which can be either on a single chain or part of dimer or oligomer complexes (21, 42). Second, the recognition motifs of their single RNA-binding domains are usually similar (Figure 2). These two aspects make it simple to evolve the sequence features in the target RNAs required for highly specific cooperative binding: a sufficient density of the simple core recognition motifs. We have recently shown that the RBP binding affinity through cooperative binding of multiple RNA-binding domains depends on the motif density on the target RNA with a Hill-like coefficient that is similar in size to the number of binding domains (37, Fig. 4D). Via di- and oligomerization of RBPs the number of cooperatively binding domains and thereby the Hill-like coefficient can be further increased, by which it is possible to distinguish between targets with, say, a core binding motif every 20 versus every

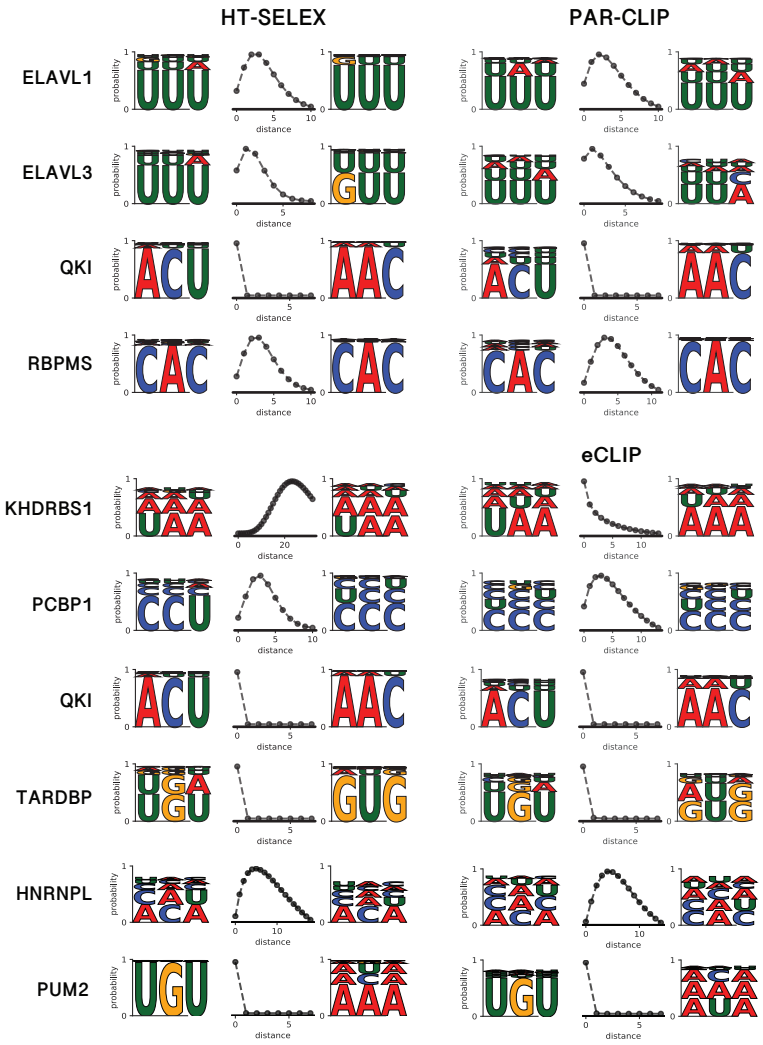


Figure 4: **Bipartite motif models learned on *in-vitro* data match their *in-vivo* counterparts** Bipartite motifs are shown for those RBPs in Figure 3 whose best replicate has a tool-averaged AUROC of at least 0.8. The models learned in-vitro and in-vivo match not only in the sequence preference but also the relative positioning of the two motif cores, with the exception of KHDRBS1, which shows a bipartite motif only in the HTR-SELEX data.

30 nucleotides (e.g 33, Fig. 3EF). An encoding of binding affinity via the density of motifs makes sense for the many RNA-binding proteins for which the precise binding sites on their target RNAs is not important to perform their function.

Mono- and dinucleotide repeats are particularly attractive as target motifs because they possess one binding site per position and per two positions, respectively. The high density of motifs gives rise to high affinities through the combinatorially many possible binding configurations of two or more RNA-binding domains. BMF takes full account of this combinatorial complexity.

A limitation of the evolutionary strategy to bind low-complexity sequences using multiple domains with near-identical motifs is the much smaller number of motifs than can be distinguished, only 64 for length-3 cores. This low number might be sufficient, however, for targeting such RBPs to their RNA targets because specificity is enhanced by compartmentalization – an RBP occurring only in the nucleus cannot bind to cytosolic mRNAs, for example. Furthermore, only a fraction of RBPs is expressed in any one cell type at any one time, in a similar way as the many transcription factors having the same binding affinities are usually expressed in different cells or at different times.

Our results agree with previous studies that reported bipartite motifs in HTR-SELEX and RBNS datasets by counting spaced *k*-mers of various linker lengths (13, 6). The motifs we report are congruent with those reported before and additionally provide a distance distribution to describe the best binding geometry. The observation that motifs are repetitive and degenerate is also consistent with previous high-throughput studies (6).

Interestingly, BMF motifs were shorter and less complex than those reported by 13. For RBPs for which Jolma et al obtained long motifs (i.e. PCBP1, PUM1, and TARDBP), longer motif cores than 3 nucleotides in BMF could not improve prediction performance in the

cross-platform benchmark. This indicates that 3-6 base long motifs would suffice in explaining the sequence specificities for the majority of RBPs.

BMF does not take RNA secondary structure into account. RNA molecules can fold onto themselves and take various tertiary structures (41). It has been shown that some RBPs at least partially identify their target RNA molecules through binding specific structural elements (14, 22). This could further narrow the search space of proteins to fewer potential binding partners and open new ways for cellular regulation. Despite ignoring structure, BMF's performance is comparable if not better than GraphProt, a tool that includes detailed modelling of secondary structure. We expect that expanding our bipartite motif model to include RNA structure could further improve its predictive power.

Overall, BMF's performance is promising in the following regards: Owing to its multi-domain binding model BMF can (1) find pairs of sequence motifs overrepresented in a sequence set, and can (2) learn the distance between the motif pairs, reflecting the best binding configurations. This information can be further used to (3) asses whether or not an RBP displays bipartite binding. We believe that looking at RNA motifs as combinations of individual low affinity interactions can improve our understanding of RNA regulation in the cell and shed a new light on how some RBPs can find their targets despite the weak sequence and structural preferences of individual domains.

Code and data availability

The HTR-SELEX data of 13 were downloaded from the European Nucleotide Archive under accession PRJEB25907 (https://www.ebi.ac.uk/ena/browser). The preprocssed eCLIP datasets were collected from the ENCODE at https://www.encodeproject.org (40). PAR-CLIP peaks were obtained from https://github.com/BIMSBbioinfo/RCAS_meta-analysis (24). BMF source code, documentation, and motif models can be found at https://github.com/-soedinglab/bipartite_motif_finder.

Acknowledgements

We thank Christian Roth for his help on AVX-optimization, web server implementation and for discussions on tool development and benchmark design. SSJ acknowledges support from the International Research School for Molecular Biology (IMPRS-MolBio). We acknowledge support by the focus program SPP2191 of the Deutsche Forschungsgemeinschaft.

References

1 Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature Biotechnol*, **33**(8), 831–838.

2 Änkö, M.-L. and Neugebauer, K. M. (2012). RNA-protein interactions in vivo: global gets specific. *Trends Biochem. Sci*, **37**(7), 255–262.

3 Ben-Bassat, I., Chor, B., and Orenstein, Y. (2018). A deep neural network approach for learning intrinsic protein-RNA binding preferences. *Bioinformatics*, **34**(17), i638–i646.

4 Calabretta, S. and Richard, S. (2015). Emerging roles of disordered sequences in RNA-binding proteins. *Trends Biochem Sci*, **40**(11), 662–672.

5 Cook, K. B., Vembu, S., Ha, K. C., Zheng, H., Laverty, K. U., Hughes, T. R., Ray, D., and Morris, Q. D. (2017). RNAcompete-S: Combined RNA sequence/structure preferences for RNA binding proteins derived from a single-step in vitro selection. *Methods*, **126**, 18–28.

6 Dominguez, D., Freese, P., Alexis, M. S., Su, A., Hochman, M., Palden, T., Bazile, C., Lambert, N. J., Van Nostrand, E. L., Pratt, G. A., *et al.* (2018). Sequence, structure, and context preferences of human RNA binding proteins. *Mol Cell*, **70**(5), 854–867.

7 Gerstberger, S., Hafner, M., and Tuschl, T. (2014). A census of human RNA-binding proteins. *Nature Rev Genet*, **15**(12), 829–845.

8 Ghanbari, M. and Ohler, U. (2020). Deep neural networks for interpreting RNA-binding protein target preferences. *Genome Res*, **30**(2), 214–226.

9 Gilbert, C. and Svejstrup, J. Q. (2006). RNA immunoprecipitation for determining RNA-protein associations in vivo. *Curr Prot Mol Biol*, **75**(1), 27–4.

10 Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano Jr, M., Jungkamp, A.-C., Munschauer, M., *et al.* (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**(1), 129–141.

11 Hartmann, H., Guthöhrlein, E. W., Siebert, M., Luehr, S., and Söding, J. (2013). P-value-based regulatory motif discovery using positional weight matrices. *Genome Res*, **23**(1), 181–194.

12 Hentze, M. W., Castello, A., Schwarzl, T., and Preiss, T. (2018). A brave new world of RNA-binding proteins. *Nature Rev Mol Cell Biol*, **19**(5), 327.

13 Jolma, A., Zhang, J., Mondragón, E., Morgunova, E., Kivioja, T., Laverty, K. U., Yin, Y., Zhu, F., Bourenkov, G., Morris, Q., *et al.* (2020). Binding specificities of human RNA-binding proteins toward structured and linear RNA sequences. *Genome Res*, **30**(7), 962–973.

14 Jones, S., Daley, D. T., Luscombe, N. M., Berman, H. M., and Thornton, J. M. (2001). Protein–RNA interactions: a structural analysis. *Nucleic Acids Res*, **29**(4), 943–954.

15 Kazan, H., Ray, D., Chan, E. T., Hughes, T. R., and Morris, Q. (2010). RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Comput Biol*, **6**(7), e1000832.

16 Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

17 Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M., and Zavolan, M. (2011). A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nature Methods*, **8**(7), 559–564.

18 König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M., and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature Struct Mol Biol*, **17**(7), 909.

19 Lambert, N., Robertson, A., Jangi, M., McGeary, S., Sharp, P. A., and Burge, C. B. (2014). RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. *Mol Cell*, **54**(5), 887–900.

20 Li, X., Kazan, H., Lipshitz, H. D., and Morris, Q. D. (2014). Finding the target sites of RNA-binding proteins. *Wiley Interdisciplinary Reviews: RNA*, **5**(1), 111–130.

21 Lunde, B. M., Moore, C., and Varani, G. (2007). RNA-binding proteins: modular design for efficient function. *Nature Rev Mol Cell Biol*, **8**(6), 479–490.

22 Mackereth, C. D. and Sattler, M. (2012). Dynamics in multi-domain protein recognition of RNA. *Curr Opin Struct Biol*, **22**(3), 287–296.

23 Maticzka, D., Lange, S. J., Costa, F., and Backofen, R. (2014). GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol*, **15**(1), 1–18.

24 Mukherjee, N., Wessels, H.-H., Lebedeva, S., Sajek, M., Ghanbari, M., Garzia, A., Munteanu, A., Yusuf, D., Farazi, T., Hoell, J. I., *et al.* (2019). Deciphering human ribonucleoprotein regulatory networks. *Nucleic Acids Res*, **47**(2), 570–581.

25 Munteanu, A., Mukherjee, N., and Ohler, U. (2018). SSMART: sequence-structure motif identification for RNA-binding proteins. *Bioinformatics*, **34**(23), 3990–3998.

26 Orenstein, Y. and Shamir, R. (2014). A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data. *Nucleic Acids Res*, **42**(8), e63–e63.

27 Ozdilek, B. A., Thompson, V. F., Ahmed, N. S., White, C. I., Batey, R. T., and Schwartz, J. C. (2017). Intrinsically disordered RGG/RG domains mediate degenerate specificity in RNA binding. *Nucleic Acids Res*, **45**(13), 7984–7996.

28 Pan, X. and Shen, H.-B. (2018). Predicting RNA–protein binding sites and motifs through combining local and global deep convolutional neural networks. *Bioinformatics*, **34**(20), 3427–3436.

29 Quinn, T. P., Nguyen, D., Nguyen, P., Gupta, S., and Venkatesh, S. (2020). Learning distance-dependent motif interactions: an interpretable CNN model of genomic events. *bioRxiv*.

30 Ray, D., Kazan, H., Cook, K. B., Weirauch, M. T., Najafabadi, H. S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., *et al.* (2013). A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, **499**(7457), 172–177.

31 Rubinstein, M., Colby, R. H., *et al.* (2003). *Polymer Physics*, volume 23. Oxford university press New York.

32 Schneider, T., Hung, L.-H., Aziz, M., Wilmen, A., Thaum, S., Wagner, J., Janowski, R., Müller, S., Schreiner, S., Friedhoff, P., *et al.* (2019). Combinatorial recognition of clustered RNA elements by the multidomain RNA-binding protein imp3. *Nature Commun*, **10**(1), 1–18.

33 Schulz, D., Schwalb, B., Kiesel, A., Baejen, C., Torkler, P., Gagneur, J., Soeding, J., and Cramer, P. (2013). Transcriptome surveillance by selective termination of noncoding RNA synthesis. *Cell*, **155**(5), 1075–1087.

34 Siebert, M. and Söding, J. (2016). Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Res*, **44**(13), 6055–6069.

35 Singh, G., Pratt, G., Yeo, G. W., and Moore, M. J. (2015). The clothes make the mRNA: past and present trends in mrnp fashion. *Ann Rev Biochem*, **84**, 325–354.

36 Singh, R. and Valcárcel, J. (2005). Building specificity with nonspecific RNA-binding proteins. *Nature Struct Mol Biol*, **12**(8), 645–653.

37 Stitzinger, S. H., Sohrabi-Jahromi, S., and Söding, J. (2021). Cooperativity boosts affinity and specificity of proteins with multiple RNA-binding domains. *bioRxiv*.

38 Stražar, M., Žitnik, M., Zupan, B., Ule, J., and Curk, T. (2016). Orthogonal matrix factorization enables integrative analysis of multiple RNA binding proteins. *Bioinformatics*, **32**(10), 1527–1535.

39 Van Nostrand, E. L., Pratt, G. A., Shishkin, A. A., Gelboin-Burkhart, C., Fang, M. Y., Sundararaman, B., Blue, S. M., Nguyen, T. B., Surka, C., Elkins, K., *et al.* (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature Methods*, **13**(6), 508–514.

40 Van Nostrand, E. L., Pratt, G. A., Yee, B. A., Wheeler, E. C., Blue, S. M., Mueller, J., Park, S. S., Garcia, K. E., Gelboin-Burkhart, C., Nguyen, T. B., *et al.* (2020). Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins. *Genome Biol*, **21**, 1–26.

41 Wan, Y., Kertesz, M., Spitale, R. C., Segal, E., and Chang, H. Y. (2011). Understanding the transcriptome through RNA structure. *Nature Rev Genet*, **12**(9), 641–655.

42 Wang, X., McLachlan, J., Zamore, P. D., and Hall, T. M. T. (2002). Modular recognition of RNA by a human pumilio-homology domain. *Cell*, **110**(4), 501–512.

43 Yang, X., Li, H., Huang, Y., and Liu, S. (2013). The dataset for protein–RNA binding affinity. *Protein Sci*, **22**(12), 1808–1811.

Supplemental Information

Thermodynamic modeling reveals widespread multivalent binding by RNA-binding proteins

Salma Sohrabi-Jahromi,[1] Johannes Söding[1,2*]

[1] Quantitative and Computational Biology, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany.

[2] Campus-Institut Data Science (CIDAS), Göttingen, Germany.

[*] Correspondence: soeding@mpibpc.mpg.de

Supplementary Methods

Parameter Optimization

We learn BMF model parameters by maximizing the likelihood function (equation 9 in manuscript). For an efficient optimization using stochastic gradient descent, we need to be able to compute the partial derivative of the likelihood with respect to model parameters (θ):

$$\frac{\partial LL(\mathbf{\Theta})}{\partial\theta} = \sum_{\mathbf{x}\in\mathcal{X}^+} \frac{1}{Z(\mathbf{x})(1-Z(\mathbf{x}))}\frac{\partial Z(\mathbf{x})}{\partial\theta} - N^+ \frac{\sum_{\mathbf{x}'\in\mathcal{X}^{\text{bg}}} p_{\text{bg}}(\mathbf{x}')Z(\mathbf{x}')^{-2}\times\frac{\partial Z(\mathbf{x}')}{\partial\theta}}{\sum_{\mathbf{x}'\in\mathcal{X}^{\text{bg}}} p_{\text{bg}}(\mathbf{x}')(1-1/Z(\mathbf{x}'))}. \tag{1}$$

We can compute the partial derivatives $\partial Z(\mathbf{x})/\partial\theta$ from the partial derivatives $\partial Z_{\text{A}}(i)/\partial\theta$ and $\partial Z_{\text{A}}(L)/\partial\theta$ according to equation 6 in the manuscript:

$$\frac{\partial Z(\mathbf{x})}{\partial\theta} = \frac{\partial Z_{\text{B}}(x, L-1)}{\partial\theta} + \sum_{i=0}^{L-1}\frac{\partial Z_{\text{A}}(x,i)}{\partial\theta}\,. \tag{2}$$

BMF parameters (θ) include binding energies of each domain to various k-mers as well as concentration parameters (S, r and p).

In the following, We define $\theta_{k,d}$ as the binding energy of domain d at k-mer k. Log-likelihood derivatives with respect to binding energies can be computed iteratively by applying the partial derivative operator on the forward algorithm of the dynamic programming (equations 1 and 2 in the manuscript):

$$\frac{\partial Z_{\text{A}}(i)}{\partial\theta_{k,d}} = c_{\text{AB}}e^{-E_A(i)}\left[\frac{\partial Z_{\text{B}}(i-k)}{\partial\theta_{k,d}} + \sum_{j=0}^{i-k}\frac{\partial Z_{\text{A}}(j)}{\partial\theta_{k,d}} - \left(Z_{\text{B}}(i-k) + \sum_{j=0}^{i-k} Z_{\text{A}}(j)\right)\frac{\partial E_A(i)}{\partial\theta_{k,d}}\right], \tag{3}$$

where

$$\frac{\partial E_A(i)}{\partial \theta_{k,d}} = \delta_{x(i),k}\ \delta_{d,A}\ , \tag{4}$$

and $\delta_{i,j}$ is the Kronecker delta of i and j. Similarly we can get the derivatives with respect to Z_{B}:

$$\begin{aligned}\frac{\partial Z_{\mathrm{B}}(i)}{\partial \theta_{k,d}} = \frac{\partial Z_{\mathrm{B}}(i-1)}{\partial \theta_{k,d}} + \sum_{j=0}^{i-k} c_{\mathrm{B}}(i-k-j)\, e^{-E_B(i)} \left(\frac{\partial Z_{\mathrm{A}}(j)}{\partial \theta_{k,d}} - Z_{\mathrm{A}}(j) \frac{\partial E_B(i)}{\partial \theta_{k,d}} \right) \\ + c_{\mathrm{AB}}\, e^{-E_B(i)} \left(\frac{\partial Z_{\mathrm{B}}(i-k)}{\partial \theta_{k,d}} - Z_{\mathrm{B}}(i-k) \frac{\partial E_B(i)}{\partial \theta_{k,d}} \right),\end{aligned} \tag{5}$$

where

$$\frac{\partial E_B(i)}{\partial \theta_{k,d}} = \delta_{x(i),k}\ \delta_{d,B}\ . \tag{6}$$

These derivatives are computed iteratively via dynamic programming similar to Z_{A} and Z_{B}. They are initialized to zero for:

$$\frac{\partial Z_{\mathrm{A}}(i)}{\partial \theta_{k,d}} = 0 \text{ for all } i \in \{0, ..., k-2\} \tag{7}$$

$$\frac{\partial Z_{\mathrm{B}}(i)}{\partial \theta_{k,d}} = 0 \text{ for all } i \in \{0, ..., k-2\}\,. \tag{8}$$

Similarly, we can derive the partial derivative in respect to the concentration parameters (θ_c):

$$\frac{\partial Z_{\mathrm{A}}(i)}{\partial \theta_c} = c_{\mathrm{AB}}\, e^{-E_A(i)} \left(\frac{\partial Z_{\mathrm{B}}(i-k)}{\partial \theta_c} + \sum_{j=0}^{i-k} \frac{\partial Z_{\mathrm{A}}(j)}{\partial \theta_c} \right)\ , \tag{9}$$

$$\begin{aligned}\frac{\partial Z_{\mathrm{B}}(i)}{\partial \theta_c} = \frac{\partial Z_{\mathrm{B}}(i-1)}{\partial \theta_c} + \sum_{j=0}^{i-k} e^{-E_B(i)} \left(\frac{\partial Z_{\mathrm{A}}(j)}{\partial \theta_c} c_{\mathrm{B}}(i-k-j) + Z_{\mathrm{A}}(j) \frac{\partial c_{\mathrm{B}}(i-k-j)}{\partial \theta_c} \right) \\ + \frac{\partial Z_{\mathrm{B}}(i-k)}{\partial \theta_c} c_{\mathrm{AB}} + Z_{\mathrm{B}}(i-k) \frac{\partial c_{\mathrm{AB}}}{\partial \theta_c}\ .\end{aligned} \tag{10}$$

The partial derivative $\partial c_{\mathrm{B}} / \partial \theta_c$ with respect to concentration parameters S, r and p are (according to equation 5 in the manuscript):

$$\frac{\partial c_{\mathrm{B}}(d)}{\partial S} = \frac{\Gamma(d+r)}{\Gamma(d+1)\Gamma(r)} p^r (1-p)^d, \tag{11}$$

$$\begin{aligned}\frac{\partial c_{\mathrm{B}}(d)}{\partial p} = S \times \left(\frac{r}{p} - \frac{d}{1-p}\right) \times \exp\big(\log\Gamma(d+r) \\ - \log\Gamma(d+1) - \log\Gamma(r) + d\log(1-p) + r\log p \big),\end{aligned} \tag{12}$$

$$\begin{aligned}\frac{\partial c_{\mathrm{B}}(d)}{\partial r} = S \times \big(\psi(d+r) + \log p - \psi(r) \big) \times \exp\big(\log\Gamma(d+r) \\ - \log\Gamma(d+1) - \log\Gamma(r) + d\log(1-p) + r\log p \big),\end{aligned} \tag{13}$$

where Γ is the gamma function and ψ is its logarithmic derivative, also known as the digamma function. Note that for numerical accuracy, we have calculated the derivatives of the $\exp(\log c_{\mathrm{B}}(d))$ function, with respect to p and r.

Overall, these equations allow us to iteratively compute for any sequence $\mathbf{x}$ the partial derivatives $\partial Z_{\mathrm{A}}(i)/\partial\theta$ and $\partial Z_{\mathrm{B}}(L)/\partial\theta$ with respect to all parameters and hence derivatives of the partition function $Z(x)$ and those of the likelihood.

The thermodynamic model contains a simplification. We assume that one of the domains (A) always binds upstream of the other domain (B) and that the binding configurations A-B and B-A do not *both* contribute appreciably to the binding probability. This seems like a very plausible assumption considering that the linkers between structural domains are usually quite short, and changing the order of binding would usually result in an impossible or much less favorable (tighter) configuration of the RNA chain.

Calculation of motif entropy

To derive the entropy for each bipartite motif model, we calculate the weighted probability for each base as

$$P_b = \frac{\sum\limits_{x\in k\text{-mers}} n_{b,x}p_x + \sum\limits_{y\in k\text{-mers}} n_{b,y}p_y}{\sum\limits_{b\in N}\left(\sum\limits_{x\in k\text{-mers}_A} n_{b,x}p_x + \sum\limits_{y\in k\text{-mers}} n_{b,y}p_y\right)}, \tag{14}$$

where N is the set of nucleotides ($\{A, C, G, U\}$). We calculate the entropy as

$$\text{Entropy} = -\sum_{b\in N} P_b \log_2 P_b \,. \tag{15}$$

To establish a baseline for the observed entropy values, we generated artificial bipartite motifs where the k-mer probabilities are taken from the observed probabilities of an experimental set but the k-mers were shuffled. We generated 10,000 such motifs and used the resulting entropy distribution as a baseline for motif complexity.

Calculation of motif repetitiveness

To quantify the degree of sequence repetitiveness in BMF models, we calculate the highest average probability of observing a repetitive 3-mer (i.e. 'AUA', 'UUU', or 'CGC') as

$$R = \max_{a,b\in N}\left(\sqrt{p_A(aba) + p_A(bab))(p_B(aba) + p_B(bab))}\right), \tag{16}$$

where N is the set of nucleotides ($\{A, C, G, U\}$), and p_A and p_B are BMF probabilities for the first and second motif core respectively. To establish a baseline for the observed repetitiveness values, we calculated this metric for 10,000 artificial bipartite motifs, generated as described above.

BMF comparison with single-occurrence motif model

To estimate the effect of considering all binding configurations and including cooperativity in BMF, we compared its cross-validated classification performance with a 6-mer motif model. We created training and test sets by splitting the HTR-SELEX data with an 80 to 20 ratio. For the 6-mer model, we calculated enrichment factors for each 6-mer in training data and scored the test sequences by the

most enriched 6-mer motif. Similarly, we trained BMF with core size 3 on the training data and used the learned models to predict the binding scores for each sequence in the test set. To compare each model's classification power, we calculated the area under the receiver operating characteristic curve (AUROC) for all RBPs in the dataset.

Supplementary figures

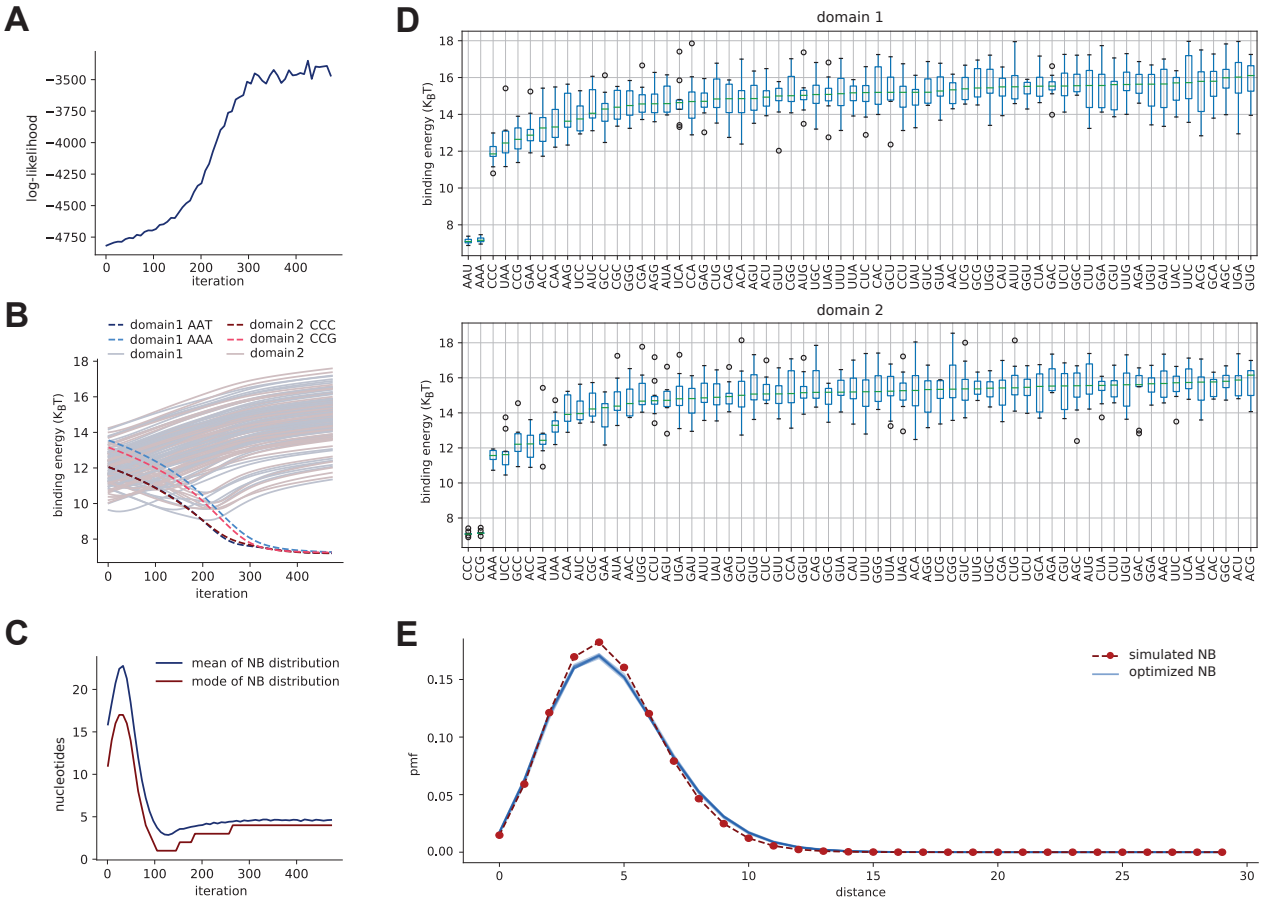


Figure S1: **BMF can reliably learn planted motifs in synthetic data.** We planted AA(A/U) followed by CC(C/G) with a distance distribution around 4 in 2000 randomly generated sequences of length 40. **(A)** Log-likelihood function increases over the iterations until reaching a plateau at the end of optimization. **(B)** The binding energies of all 3-mers are shown over BMF iterations for both binding domains. The 3-mers representing the implanted motifs are shown with brighter blue (first domain) and red (second domain)dash lines. The final values retrieved after optimization is notably lower for the highlighted 3-mers. **(C)** The mean (in blue) and mode (in red) of the NB distribution is shown over BMF's optimization iterations. The correct distance distribution is found when the LL and the energy parameters reach their plateaus. **(D)**, and **E** The distribution of final BMF parameters upon 10 random parameter initializations and subsequent optimization. Regardless of the choice of initial parameter values, BMF ends in the same optimum point in the parameter landscape.

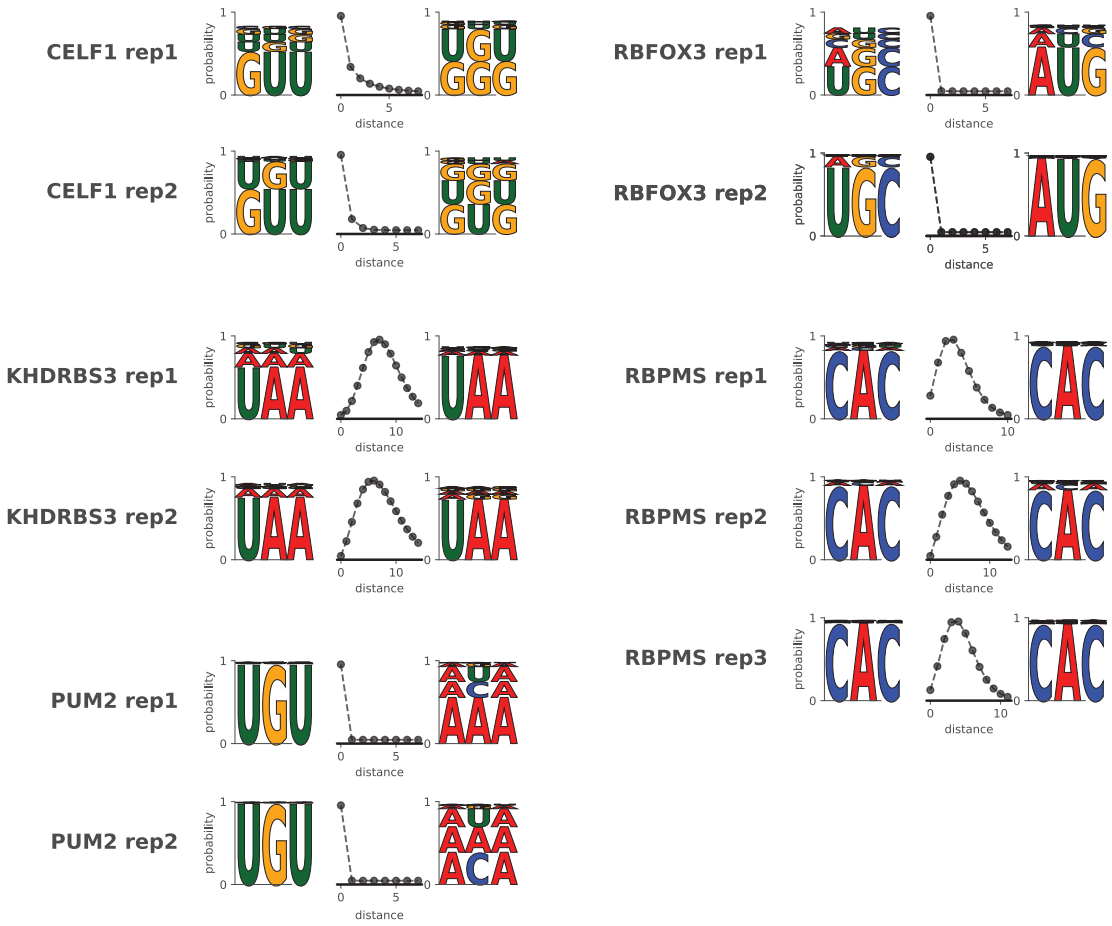


Figure S2: **Experimental HTR-SELEX replicates generate the same bipartite motif models**. Biparite binding models are shown for factors in Figure 2 for which an experimental replicate was available. The models generated for all HTR-SELEX datasets can be found in BMF GitHub repository: `https://github.com/soedinglab/bipartite_motif_finder/blob/main/data/HTRSELEX_motifs.pdf`.

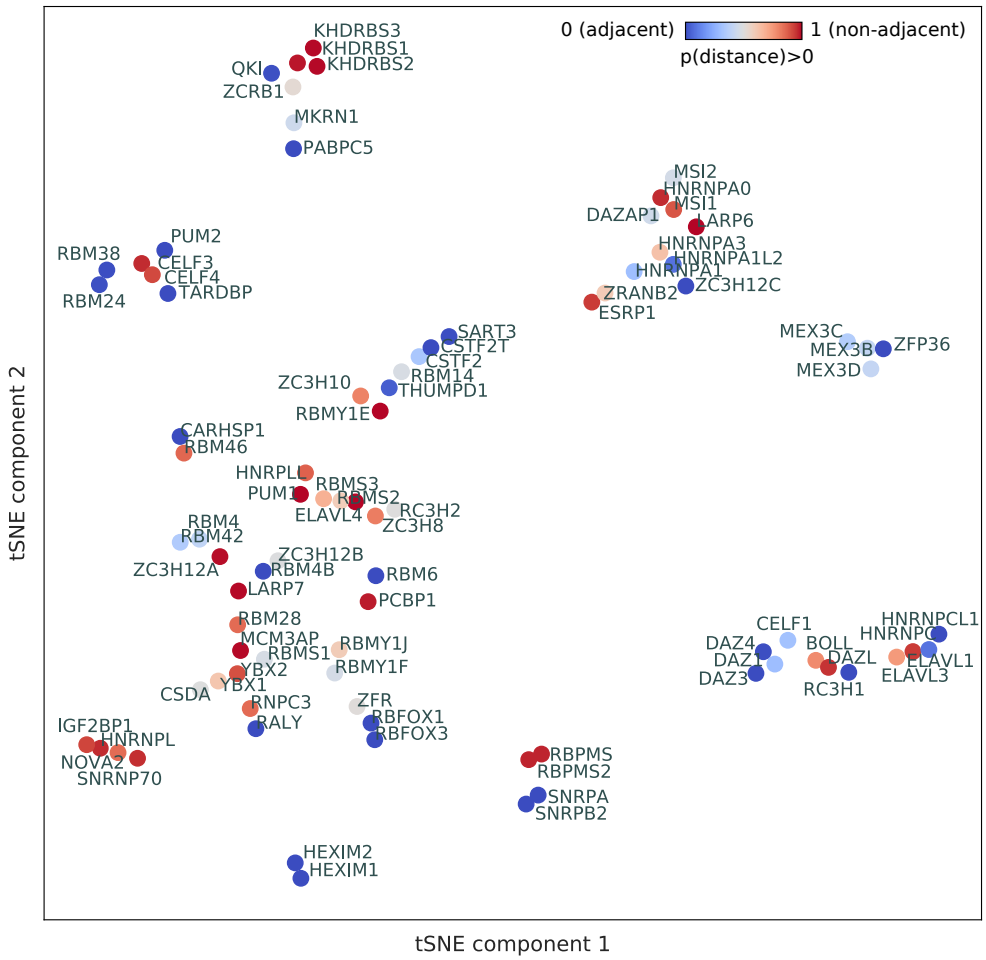


Figure S3: **RBPs in the same family have similar BMF motifs**. RBPs are clusters according to their sequence identity measured as pairwise Pearson correlation between 3-mer probabilities. Two dimensional embedding is generated via tSNE [1]. RBPs are color-coded based on the domain positioning in the NB models, as in Figure 2B, with adjacent cores colored in blue and bipartite motifs in red.

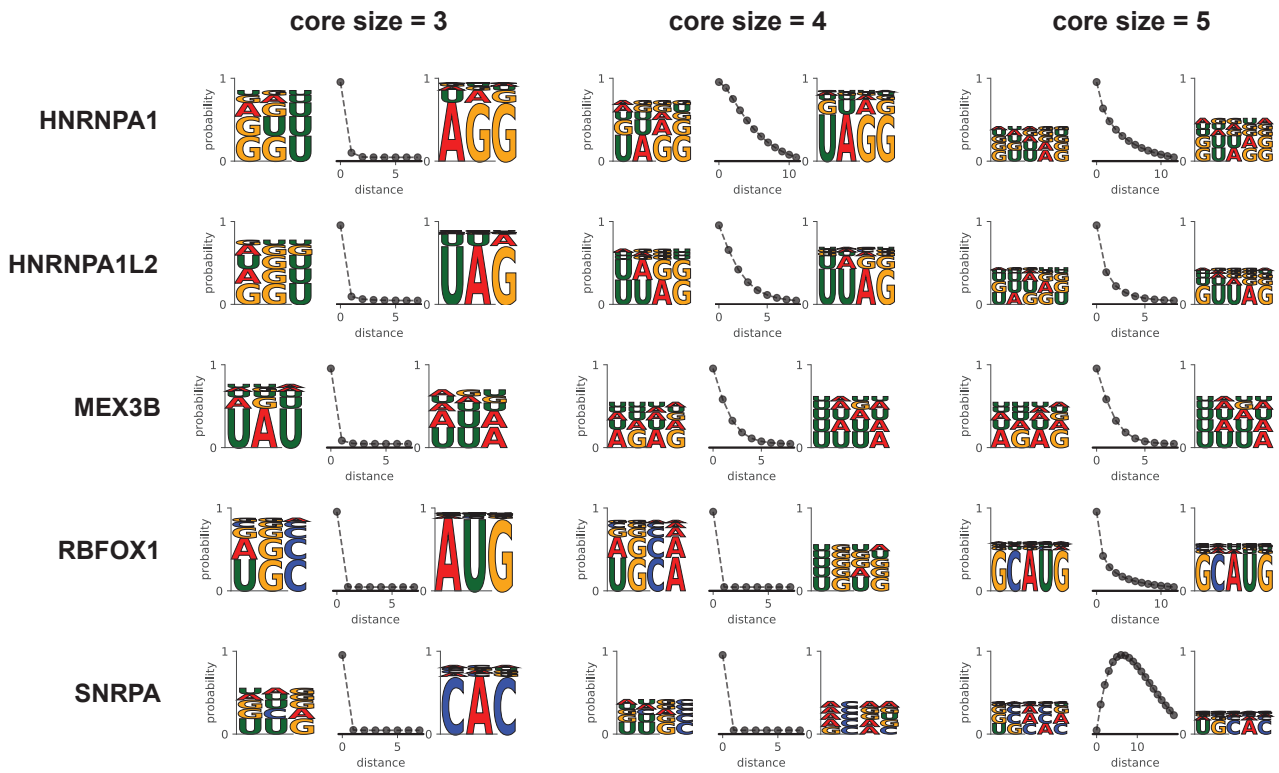


Figure S4: **Bipartite binding behaviour can arise when biulding longer sequence models**. Some RBPs in the HT-SELEX dataset have adjacent cores when building BMF models with 3-mers, but show bipartite binding for 4-mer and\or 5-mer models.

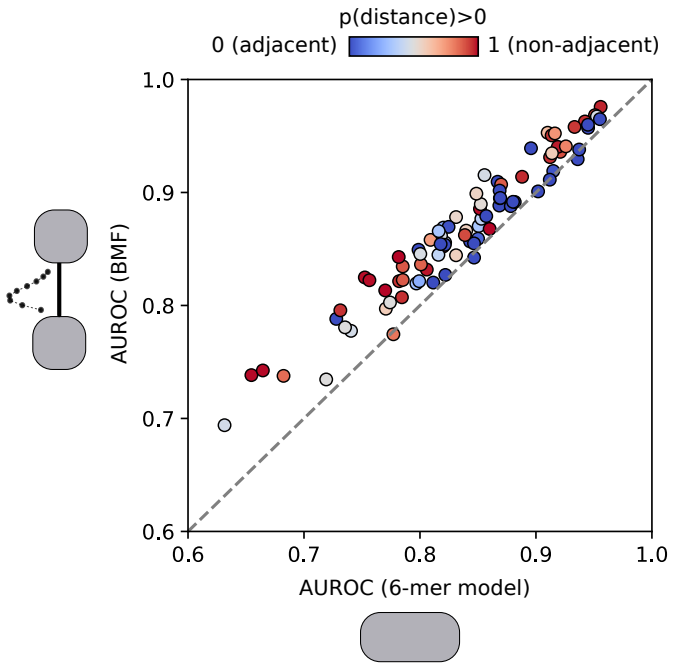


Figure S5: **Incorporating cooperativity and multivalency boosts performance of RBP binding models.** AUROC values are calculated by predicting binding sites in held-out sequences of HTR-SELEX datasets (80%-20% split for training and testing). BMF with core size 3 is compared to a single-occurence per sequence 6-mer model. RBPs are color-coded based on the domain positioning in the NB models, as in Figure 2B, with adjacent cores colored in blue and bipartite motifs in red.

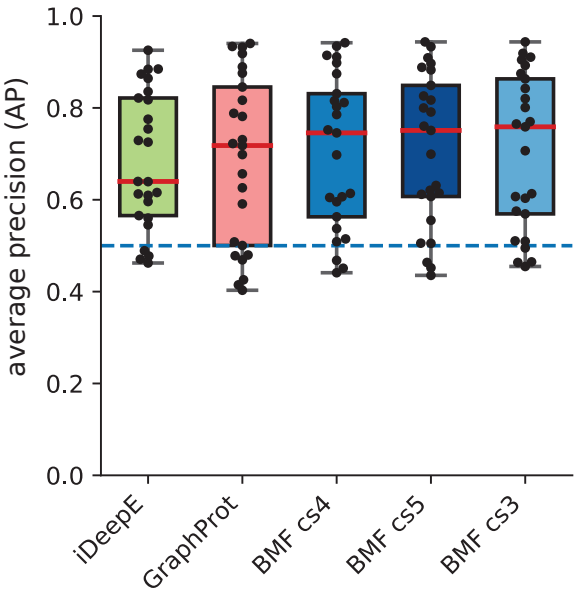


Figure S6: **Average precision (AP) scores for iDeepE, GraphProt and BMF with motif sizes ranging from 3 to 5.** We used BMF, iDeepE, and GraphProt to identify eCLIP and PAR-CLIP RBP binding sites based on the models trained on HTR-SELEX datasets. The tools are sorted based on their median AP scores (red lines). The AP score for each RBP dataset is shown with a black dot.

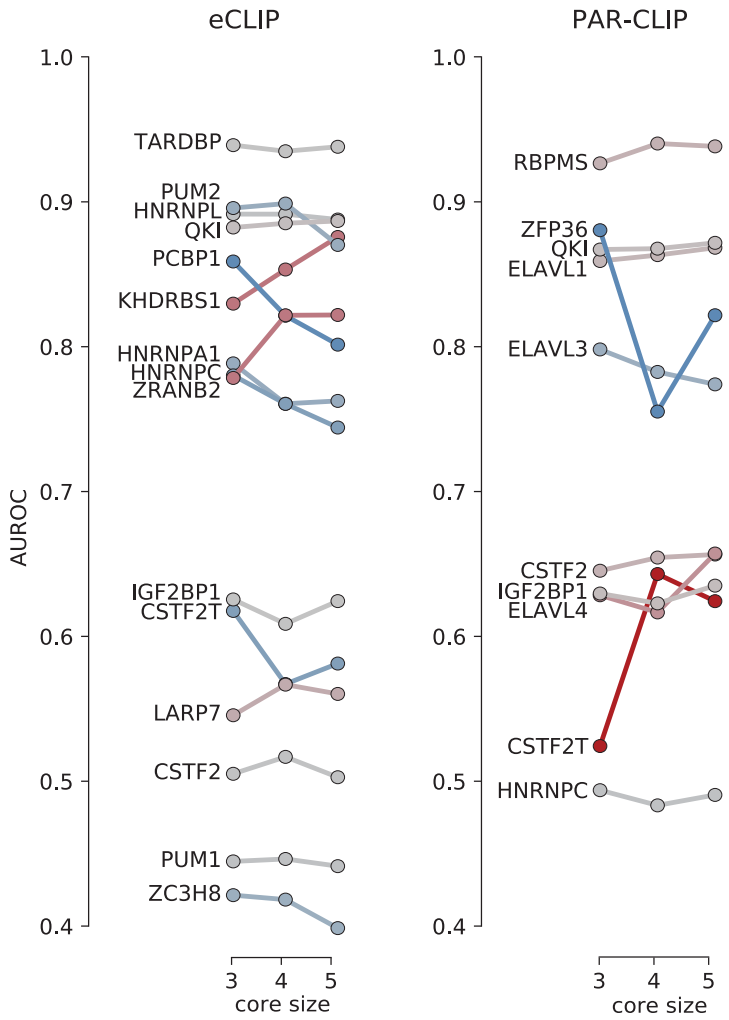
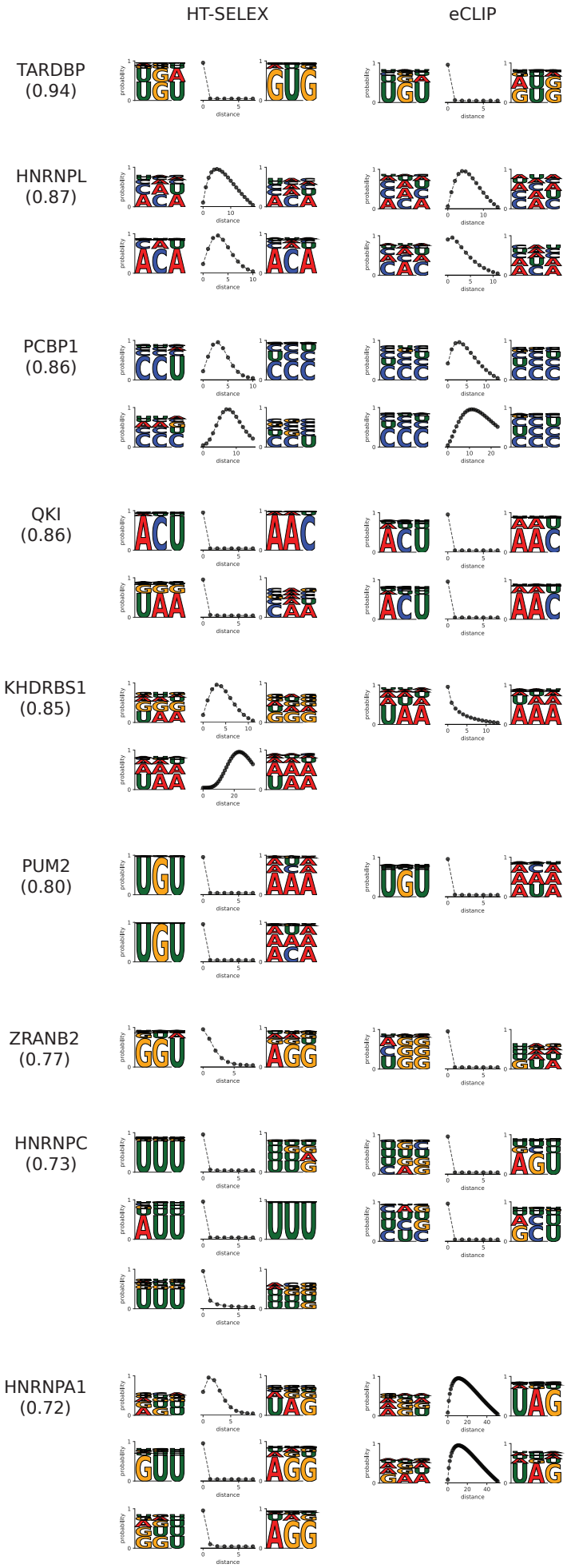


Figure S7: **Comparison of cross-platform AUROC values for BMF models with core sizes 3 to 5.** An increase in AUROC with increasing motif length is marked with red and a decrease with blue.


HT-SELEX
eCLIP
TARDBP (0.94)
HNRNPL (0.87)
PCBP1 (0.86)
QKI (0.86)
KHDRBS1 (0.85)
PUM2 (0.80)
ZRANB2 (0.77)
HNRNPC (0.73)
HNRNPA1 (0.72)
probability
distance

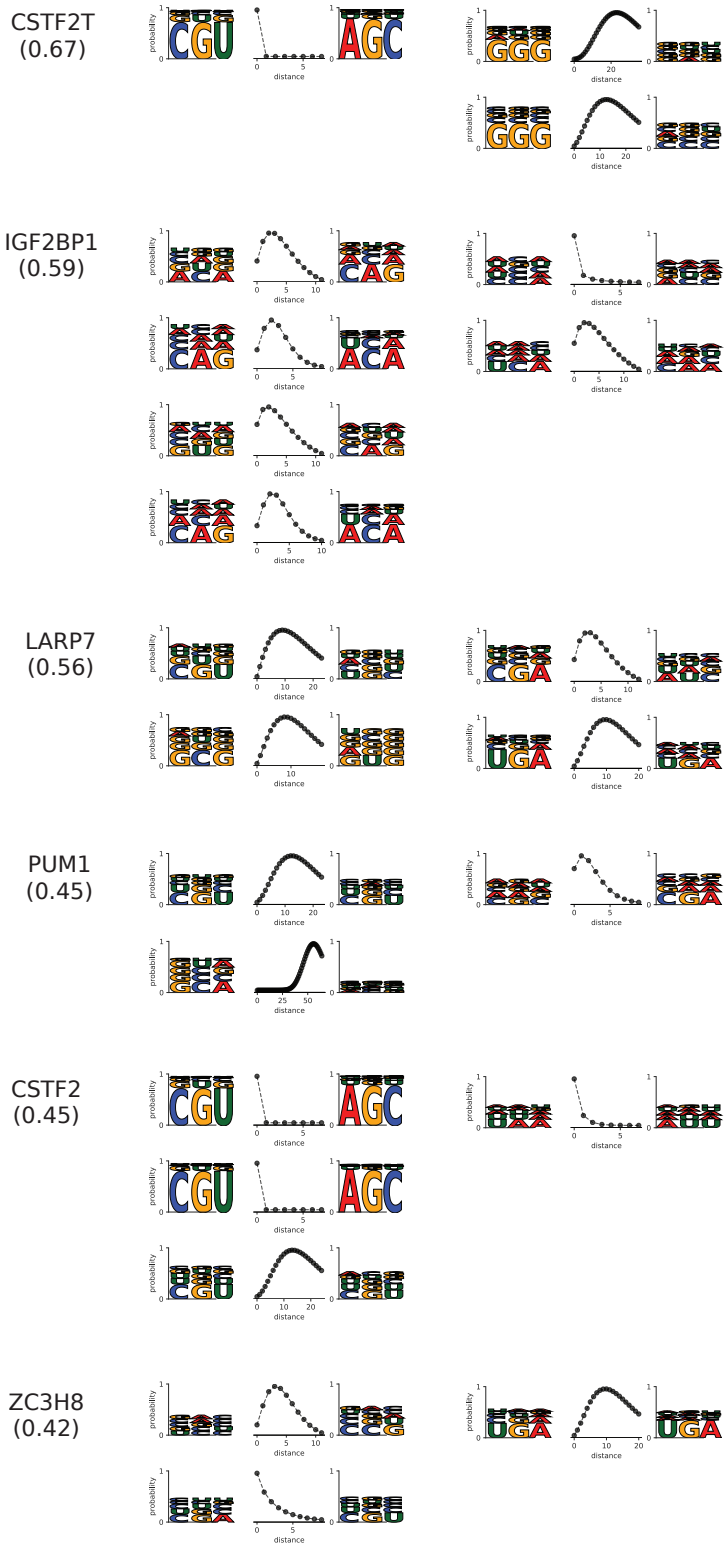
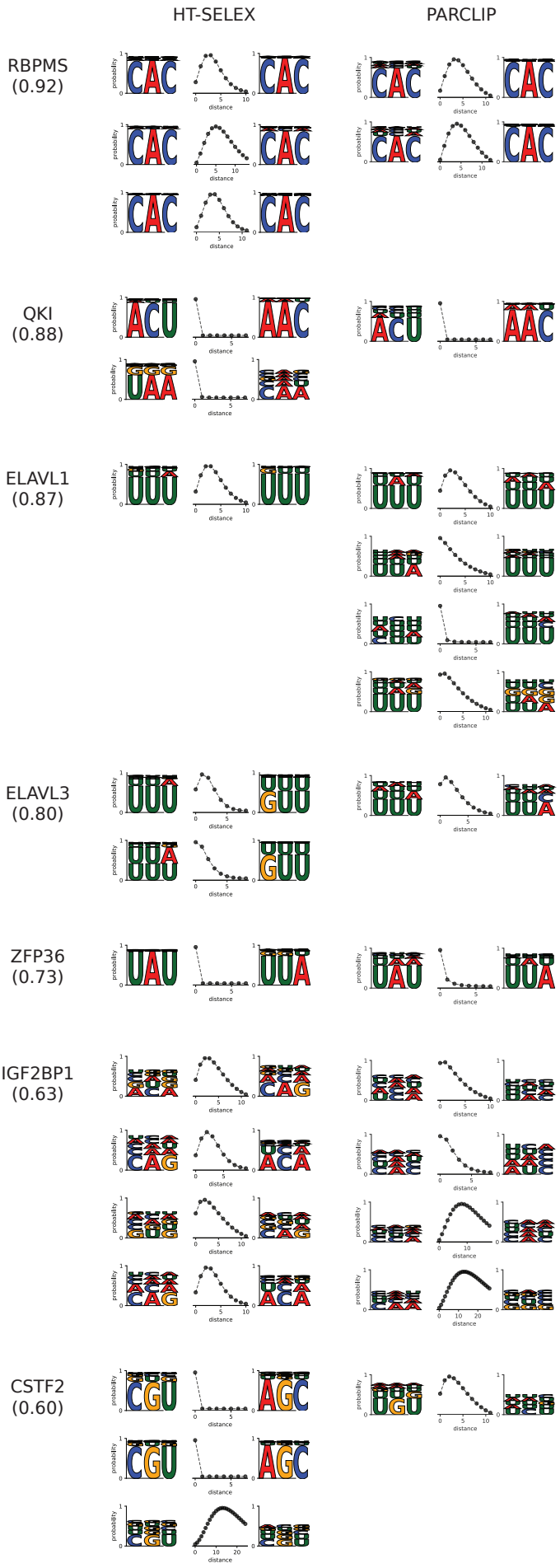


Figure S8: **Comparison of HTR-SELEX and eCLIP BMF logos**. BMF logos are sorted according to their cross-platform AUROC performance (shown in parenthesis), and is an average between BMF (with core size 3), Graphprot, and iDeepE. BMF logos were generated for all available replicates of each experimental technique.


HT-SELEX
PARCLIP
RBPMS
(0.92)
QKI
(0.88)
ELAVL1
(0.87)
ELAVL3
(0.80)
ZFP36
(0.73)
IGF2BP1
(0.63)
CSTF2
(0.60)
probability
distance

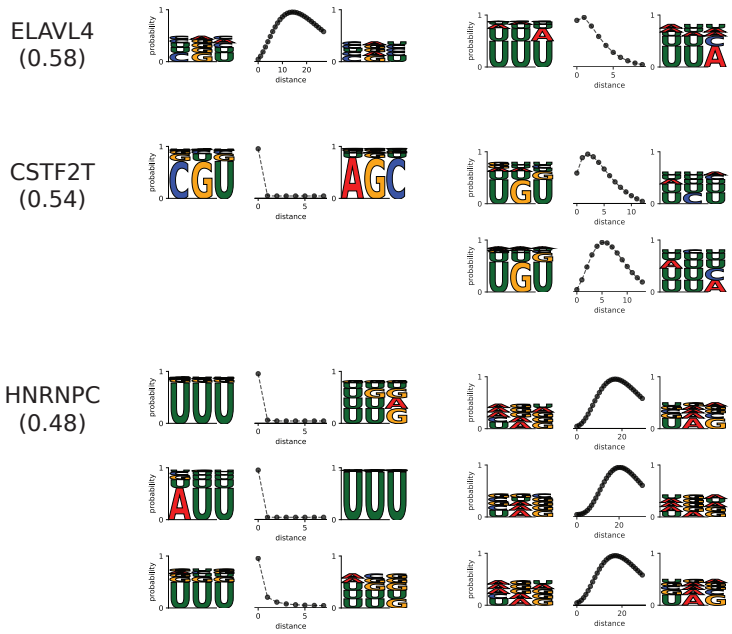


Figure S9: **Comparison of HTR-SELEX and PAR-CLIP BMF logos**. BMF logos are sorted according to their cross-platform AUROC performance (shown in parenthesis), and is an average between BMF (with core size 3), Graphprot, and iDeepE. BMF logos were generated for all available replicates of each experimental technique.

References

1 Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, **9**(Nov), 2579–2605.

4 Further contributions

4.1 Cooperativity boosts affinity and specificity of proteins with multiple RNA-binding domains.

Publication:

"Cooperativity boosts affinity and specificity of proteins with multiple RNA-binding domains."

S.H. Stitzinger, **S. Sohrabi-Jahromi**, J. Söding †

(†) corresponding author

bioRxiv (2021).

4.1.1 Manuscript abstract

Numerous cellular processes rely on the binding of proteins with high affinity to specific sets of RNAs. Yet most RNA binding domains display low specificity and affinity, to the extent that for most RNA-binding domains, the enrichment of the best binding motif measured by high-throughput RNA SELEX or RNA bind-n-seq is usually below 10-fold, dramatically lower than that of DNA-binding domains. Here, we develop a thermodynamic model to predict the binding affinity for proteins with any number of RNA-binding domains given the affinities of their isolated domains. For the four proteins in which affinities for individual domains have been measured the model predictions are in good agreement with experimental values. The model gives insight into how proteins with multiple RNA-binding domains can reach affinities and specificities orders of magnitude higher than their individual domains. Our results contribute towards resolving the conundrum of missing specificity and affinity of RNA binding proteins and underscore the need for bioinformatic methods that can learn models for multi-domain RNA binding proteins from high-throughput *in-vitro* and *in-vivo* experiments.

4.1.2 Author contributions

S.H. Stitzinger (SHS) implemented the algorithms and created all the visualizations. J. Söding (JS) conceptualized the idea. **S. Sohrabi-Jahromi (SSJ)** and JS supervised research. SHS, **SSJ**, and JS wrote the manuscript.

4.2 Mechanisms for active regulation of biomolecular condensates

Publication:

"Mechanisms for active regulation of biomolecular condensates"

J. Söding†, D. Zwicker, **S. Sohrabi-Jahromi**, M. Boehning, J. Kirschbaum

(†) corresponding author

Trends in Cell Biology, 30 (2020): 4-14.

4.2.1 Manuscript abstract

Liquid-liquid phase separation is a key organizational principle in eukaryotic cells, on par with intracellular membranes. It allows cells to concentrate specific proteins into condensates, increasing reaction rates and achieving switch-like regulation. However, it is unclear how cells trigger condensate formation or dissolution and regulate their sizes. We predict from first principles two mechanisms of active regulation by post-translational modifications such as phosphorylation: In enrichment-inhibition, the regulating modifying enzyme enriches in condensates and the modifications of proteins inhibit their interactions. Stress granules, Cajal bodies, P granules, splicing speckles, and synapsin condensates obey this model. In localization-induction, condensates form around an immobilized modifying enzyme, whose modifications strengthen protein interactions. Spatially targeted condensates formed during transmembrane signaling, microtubule assembly, and actin polymerization conform to this model. The two models make testable predictions that can guide studies into the many emerging roles of biomolecular condensates.

4.2.2 Author contributions

J. Söding (JS) and D. Zwicker (DZ) initiated the study. JS designed and prepared main figures and wrote the manuscript with input from all authors. **S. Sohrabi-Jahromi** drafted introduction and evidence supporting localization-induction. M. Boehning drafted evidence supporting enrichment-inhibition. D.Z. and J. Kirschbaum contributed to theoretical modelling.

4.3 High-throughput screen and modeling of transcription activation domains

Publication:

> "A high-throughput screen for transcription activation domains reveals their sequence features and permits prediction by deep learning."
>
> A. Erijman, L. Kozlowski, **S. Sohrabi-Jahromi**, J. Fishburn, L. Warfield, J. Schreiber, W.S. Noble, J. Söding†, S. Hahn†
>
> (†) corresponding author
>
> *Molecular Cell*, 78 (2020): 890-902.e6.

4.3.1 Manuscript abstract

Acidic transcription activation domains (ADs) are encoded by a wide range of seemingly unrelated amino acid sequences, making it difficult to recognize features that promote their dynamic behavior, "fuzzy" interactions, and target specificity. We screened a large set of random 30-mer peptides for AD function in yeast and trained a deep neural network (ADpred) on the AD-positive and -negative sequences. ADpred identifies known acidic ADs within transcription factors and accurately predicts the consequences of mutations. Our work reveals that strong acidic ADs contain multiple clusters of hydrophobic residues near acidic side chains, explaining why ADs often have a biased amino acid composition. ADs likely use a binding mechanism similar to avidity where a minimum number of weak dynamic interactions are required between activator and target to generate biologically relevant affinity and in vivo function. This mechanism explains the basis for fuzzy binding observed between acidic ADs and targets.

4.3.2 Author contributions

J. Söding (JS) and S. Hahn (SH) conceived the project. A. Erijman (AE), L. Warfield (LW), L. Kozlowski (LK), JS, and SH designed the experiments. AE, LW, and J. Fishburn performed the wet lab work. AE, LK, **S. Sohrabi-Jahromi (SSJ)**, and J. Schreiber performed computational analysis. In particular, **SSJ** performed exploratory analysis to understand what the deep network has learnt, exploring the differences in amino acid composition and examining the effect of *in silico* mutations on well studied activation domains (contributing to the underlying analyses in Figure 2A, 4A, and 6C). AE, SH, and JS. wrote the manuscript. All authors edited and approved the manuscript.

5 Discussion and outlook

RBPs regulate various stages of RNA processing from RNA transcription to RNA maturation, localization, translation, and finally degradation. Conversely, the interactions with RNA molecules can in turn regulate the fate of RBPs. To ensure the availability of mature mRNAs at the right place, at the right time, it is therefore crucial to tightly regulate RNA-protein interactions temporally and through modulating their target specificity. While this coordination requires specific targeting of RNAs by their regulators, RBPs display preferences in binding short and degenerate RNA sequences (∼3-5 bases), which alone is not specific enough to limit the search space of RBPs. Therefore, we were lacking a quantitative understanding of how RBPs achieve high affinity and specificity through their low-affinity RBDs when I started my doctoral research.

In this work, we have addressed multiple aspects of specific RNA-protein interactions. **(1)** By studying RBPs involved in the eukaryotic RNA degradation pathway, we have shown that degradation complexes, and to some extent their RBP constituents, exhibit preferences for distinct classes of transcripts and often bind preferentially to particular locations across mRNAs. We could highlight key differences in substrate specificity between the 3′ and 5′ RNA degradation machinery and propose new functions for RNA degradation proteins based on their co-binding behaviour with other RNA processing factors (Sohrabi-Jahromi et al., 2019). **(2)** We introduced BMF, a computational tool based on thermodynamic modelling of bivalent RNA-protein interactions. BMF can quantitatively model and learn bipartite binding preferences in bound sequences of an RBP. Applying BMF on a HTR-SELEX dataset of 86 RBPs showed evidence of widespread bipartite binding with a preferential linker length between the two binding sites. BMF can predict RBP binding in the cellular context and its prediction power is competitive compared with existing tools (Sohrabi-Jahromi and Söding, 2021). Furthermore, we took a quantitative thermodynamic approach to predict the binding affinity for multi-domain RBPs, given the affinities of each individual binding domain. We show that this thermodynamic model can predict dissociation constants of multi-domain RBPs by comparing its estimations with experimental measurements. Finally, quantitative simulations based on this model demonstrate how multi-domain or oligomerized RBPs can reach affinities and specificities orders of magnitude higher than their individual domains (Stitzinger et al., 2021).

In the next sections, I will discuss our contributions with regard to each of the mentioned aspects, summarize their limitations, and propose new challenges that could advance our understanding of RBP-RNA interactions.

5.1 Transcriptome maps of general eukaryotic RNA degradation factors

In the first part of this thesis, I introduced transcriptome-wide maps of RBPs involved in the yeast RNA degradation pathway (Sohrabi-Jahromi et al., 2019). In-depth bioinformatic analysis of this dataset and its cross-correlation with previously published RBP interactomes highlighted key differences in function

and specificity among the studied proteins. We also revealed groups of RBPs that show similar RNA binding patterns and could therefore have functional associations. The occupancy patterns retrieved are in line with previous studies and additionally display several unexpected observations that merit further investigation.

Distinct binding to various RNA classes. The Pan2/Pan3 deadenylation complex and subunits of the Ccr4/Not complex showed strong binding to major classes of ncRNAs: rRNAs, tRNAs, snRNAs, and snoRNAs as well as mRNAs. This is consistent with the role of deadenylation complexes in ncRNA maturation, as well as with binding spliced, mature mRNAs at their 3′ end to shorten the poly(A) end of the transcripts (Azzouz et al., 2009; Miller and Reese, 2012; Laribee et al., 2015). However, the catalytic subunit Ccr4 displayed a distinct binding profile, with its cross-link sites enriched in introns, NUTs, and the 5′ end of mRNA molecules. This indicates that Ccr4 and Pop2 of the Ccr4/Not complex may be playing distinct roles by processing poly(A) tails of different RNA classes. This is supported by the observation that Ccr4 is also involved in transcription elongation and its regulation (Kruk et al., 2011; Reese, 2013). In line with this, our co-occupancy maps of RBPs group Ccr4 together with transcription elongation factors (Figure 5). The decapping complex on the other hand shows a stronger binding to cryptic Pol II transcripts, particularly SUTs, as well as to the 3′ end of transcripts. The binding behavior closely matches that of Xrn1, in line with its role in degrading decapped transcripts (Dendooven et al., 2020; Braun et al., 2012).

Differences between 5′ and 3′ end decay pathways. Overall the 5′ (decapping and Xrn1) and 3′ (deadenylation and exosome) degradation machineries specifically target distinct transcript classes, suggesting that RNA regulatory pathways invoke transcript decay via different routes (Figure 2 and 3). However, their distinct binding patterns across mRNAs is counter intuitive: the 5′ machinery is enriched at the poly(A) side while the 3′ degradation machinery is enriched around the cap region. We propose two plausible explanations to this puzzling observation. (1) Since the cross-link sites reflect snap-shots of RNA-protein interactions in the cell, a lower pace of exonucleation at the opposite side of the RNA molecule can result in binding enrichment of the slowing exonuclease. (2) It is also possible that this enrichment is due to the formation of closed mRNP loops and a cross-linkability bias in the opposite unbound mRNA end.

Characterization of the general nuclear surveillance and preprocessing machinery. Comparing the binding profiles with respect to cryptic Pol II transcripts indicated the involvement of TRAMP and exosome (both Rrp44 and Rrp6 exonucleases) in the Nrd1/Nab3 mediated degradation of cryptic Pol II transcripts and snoRNA processing (Figure 4, 5, and Figure 3—figure supplement 2). Numerous studies have shown the involvement of TRAMP/exosome machineries in transcription termination and provided mechanistic insights into how the Nrd1/Nab3 complex interacts with the degradation system (Fox et al., 2015; Fasken et al., 2015; Porrua and Libri, 2015; Schmid and Jensen, 2019). We additionally demonstrated that, out of the two TRAMP complexes, TRAMP4 (Mtr4, Air2, and Trf4) is the one responsible for Nrd1-mediated RNA degradation. This is consistent with the observation that Nrd1's interaction domain (which binds Pol II CTD) can also recognize a CTD mimic region in Trf4 and subsequently stimulate the polyadenylation activity of TRAMP (Tudek et al., 2014). Together these results imply that the TRAMP complexes serve partially non-overlapping functions in the nucleus.

Degradation of translationally inefficient mRNAs. Previous studies have suggested that Dhh1 may be involved in the recruitment of the degradation machinery to translationally slow mRNAs (Rad-

hakrishnan et al., 2016; Sweet et al., 2012). We show for the first time in a large scale mapping of mRNAs that the decapping complex as well as Xrn1 are significantly enriched on mRNAs with a low average codon optimality, in comparison to exosome and deadenylation complexes, which show the opposite trend (Figure 6). This implies that translation difficulties mostly result in the recruitment of the 5′ degradation RNA system, and is consistent with a later study that finds XRN1-dependent 5′ decay as the main determinant of RNA half-life in mammalian cells (Tuck et al., 2020).

Possible limitations. It is important to note that the PAR-CLIP data capture RBP occupancy on the transcripts and do not directly reveal protein function. Nonetheless, we hope that the associations observed can guide future studies and inspire rigorous and detailed biochemical assays to further characterize degradation factors. A limitation of PAR-CLIP for studying mRNAs is that poly(A)-binding events cannot be captured as the recovered sequences would not be mapped to the genome. It is therefore noteworthy to consider this when interpreting the binding profiles of the deadenylation complexes. Another challenge when working with PAR-CLIP data is that some transcriptomic sequences cross-link at a higher rate, which can create biases in the data (Friedersdorf and Keene, 2014). We reduce these biases in our analyses by taking transcriptome-wide approaches that characterizes proteins through comparing their binding profile with those obtained from other RBPs.

Taken together, this study contributes to our understanding of eukaryotic RNA degradation in several ways. **(1)** We characterize the differences in substrate specificity between the 5′ and 3′ degradation machineries. **(2)** We identify a general nuclear surveillance machinery consisting of TRAMP4, exosome, and Nrd1/Nab3 complexes, responsible for targeting aberrant nuclear RNAs as well as preprocessing snoRNAs. **(3)** We find that the decapping complex is mostly recruited only upon RNA degradation, while other decay factors apparently already associate with mRNAs earlier for their surveillance. **(4)** Our extensive RBP interactome data can provide a resource for molecular biologists studying RNA degradation, guiding future experiments.

5.2 Thermodynamic modeling of multivalent binding by RBPs

In the second part of this thesis, I introduced BMF for *de novo* discovery of bipartite motifs in RNA-protein interaction data (Sohrabi-Jahromi and Söding, 2021). To the best of our knowledge BMF is the first approach that adopts a full thermodynamic viewpoint for RNA motif identification. Instead of considering only the best binding configuration, BMF aggregates contributions from all possible binding configurations on the RNA sequence. This facilitates discovering motifs in repetitive and degenerate RNA sequences such as mRNA UTRs. We applied BMF on a HTR-SELEX dataset of 86 human RBPs (Jolma et al., 2020) and showed widespread bipartite binding with a factor-dependent preferred gap length between the motif cores. These results demonstrate the importance of multivalent binding for RBPs and contribute to our understanding of the mRNP code underlying mRNA regulation.

Differences between TF and RBP targeting and their implications. *In silico* RNA motif discovery resembles TF motif discovery as both problems aim at the identification of over-represented sequence features to explain the specific targeting by a certain protein of interest. This has resulted in the initial repurposing of many genomic motif discovery tools for *de novo* RNA motif discovery (such as Frith et al., 2008; Bailey et al., 2015; Alipanahi et al., 2015). However, RBP binding fundamentally differs from TF targeting in a number of ways that merit a careful consideration when designing computational motif detection tools. Transcription regulation relies on the binding of TFs to specific regulatory elements

around gene promoters. To this end, TFs typically read out DNA stretches spanning 6-12 base pairs (Lambert et al., 2018). RBPs on the other hand mostly regulate a large number of target RNA molecules and their binding is dynamically regulated through other binding partners that facilitate or inhibit their binding to RNA (Sternburg and Karginov, 2020). For many RBPs the precise binding locations on their target RNA molecules are not important for performing their function. For instance, proteins that bind sequence elements on mRNAs to transport them across the cytoskeleton could identify any part of the RNA molecule.

Another feature of RBPs is their high modularity, with multiple domains binding adjacent or spaced RNA fragments in a semi-flexible RNA chain (Lunde et al., 2007). RBPs therefore achieve their dynamically regulated targeting by identifying short and often degenerate sequences with each of their domains (Dominguez et al., 2018). This allows the binding of many RBPs to repetitive and degenerate RNA UTRs through cooperative effects. Their specificity can be boosted through binding preferences for certain RNA structures (Li et al., 2014), getting help from an associated small RNA (Djuranovic et al., 2011; Bartel, 2018), or interactions with other RBPs (Sternburg and Karginov, 2020; Müller-McNicoll and Neugebauer, 2013). The binding selectivity is further boosted through compartmentalization. Higher concentrations of RBPs in P-bodies can for example enhance new RBP-RNA interactions. Similarly, sequestering one of the interaction partners in the nucleus or a membraneless compartment can prevent their interaction (Hubstenberger et al., 2017; Mittag and Parker, 2018). Out of these features, RNA secondary structure has received the most attention as many recent RNA motif discovery tools utilize the RNA structure as input data. (Pan et al., 2018; Maticzka et al., 2014; Budach and Marsico, 2018; Zhang et al., 2016; Ben-Bassat et al., 2018; Su et al., 2019; Deng et al., 2020). BMF complements such approaches by incorporating multivalency, searching for pairs of short sequence patterns enriched in bound RNA fragments.

Previous reports on bivalent binding. Spaced k-mer approaches have previously suggested bipartite binding modes for about one third of RBPs in HTR-SELEX and RBNS datasets (Dominguez et al., 2018; Schneider et al., 2019; Jolma et al., 2020). BMF finds bipartite binding for half of the 78 studied RBPs. In these cases, the motif cores match in their sequence preference and the motifs have low complexity and high repetitiveness (Figure 2). The low complexity and repetitive nature of RNA motifs has been described before (Dominguez et al., 2018). The identification of low complexity sequences produces multiple binding surfaces on repetitive RNA sequences which allows RBPs to interact with higher affinity. BMF's motif model takes this combinatorial complexity into account. Overall, the models learned by BMF match previously reported motifs while providing additional information on the distance preference of the motif cores, capturing the optimal geometry of the protein binding sites.

The choice of the experimental dataset for bipartite motif discovery. *In-vitro* datasets are advantageous for motif discovery purposes as they are free of cellular complications such as effects of interactions with protein cofactors, non-specific background binding, and most protocol-induced sequence biases (Friedersdorf and Keene, 2014; Kishore et al., 2011; Dominguez et al., 2018). Moreover, the availability of a large pool of random RNA oligomers ensures a sufficiently large selection pool to discover RBP binding preferences in comparison to *in-vivo* data that are limited by the non-random transcriptome composition. In order to capture bipartite binding modes in such datasets, the RNA oligomers would need to be sufficiently long to accommodate the two motif cores as well as their linker sequence. The HTR-SELEX dataset by Jolma et al. is the only *in-vitro* large-scale dataset available that uses longer oligomers (40 nucleotides) in its selection process, and hence was used here to discover bipartite motifs.

Cross-platform validation. We introduce a cross-platform benchmark to evaluate the quality of BMF predictions. As noted in section 1.3.2, highly parametric motif models can learn biases in experimental datasets to distinguish bound from unbound RNA fragments (Ghanbari and Ohler, 2020; Kishore et al., 2011; Orenstein and Shamir, 2014). We therefore evaluate the performance of BMF in predicting *in vivo* binding sites based on models trained on *in vitro* data. To establish a baseline, we similarly gauged the prediction accuracy of the frequently used and highly parametric motif models GraphProt and iDeepE. GraphProt uses sequence and structural information and relies on support-vector machines, while iDeepE is based on deep learning (Maticzka et al., 2014; Pan and Shen, 2018). Overall, BMF can predict *in vivo* binding sites with competitive accuracy (Figure 3). This could be due to the over-fitting of more complex models to experimental artifacts that prevent them from generalizing across platforms. However, GraphProt and iDeepE excelled at predicting new binding sites for few proteins with long sequence preferences such as CSTF2T. In such cases, longer BMF models are needed to best describe the binding motifs (Figure S7).

Possible limitations. While BMF models show promising accuracy in predicting new binding sites, they do not take the RNA structure into account. Numerous studies have shown preferential binding to certain RNA structure elements or a mere binding preference towards single-stranded accessible parts of the RNA molecule (Li et al., 2014; Dominguez et al., 2018; Jolma et al., 2020). Building hybrid models of sequence and structure has therefore improved the performance of RNA motif models (Pan et al., 2018; Maticzka et al., 2014; Budach and Marsico, 2018; Zhang et al., 2016; Ben-Bassat et al., 2018; Su et al., 2019; Deng et al., 2020). We expect that incorporating the RNA structure can further improve the accuracy of BMF models. Another assumption made by BMF is that the two domains always bind in the same order to the RNA sequence and that their binding cannot be swapped. This allows us to more easily enumerate all possible binding configurations in the learning phase. However, this simplifying assumption will be justified for the vast majority of factors as one of the two orders of binding will usually lead to a much less favorable configuration due to the spatial constraints and the need for a flexible and long peptide linker between the two domains.

Thermodynamic modelling to estimate RBP dissociation constants. We further show the impact of multi-domain binding by modeling the protein RNA binding kinetics for RBPs with any number of binding sites (Stitzinger et al., 2021). We model protein and RNA linkers connecting the binding sites as flexible chains and estimate total dissociation constants (K_ds) based on K_ds derived for individual domains. This model proves promising based on its accuracy in predicting RBP K_ds, and additionally provides interesting insights on how cooperative binding can result in affinities and specificities that are orders of magnitude higher than those achieved by single domains. The kinetic simulations show that small changes in motif density can significantly boost the binding probability of multi-domain RBPs. These results together with the motif models found by BMF indicate that binding affinities may be encoded in the low-complexity RNA sequences through small variations in the number of potential binding sites.

Overall, we contribute to a better understanding of RNA-protein interactions in the following regards. **(1)** We introduce the first tool to incorporate cooperativity in motif discovery enabling detection of bipartite motifs as well as the distance preferences between the motif cores. **(2)** We perform in depth analysis of binding motifs learned from 78 RBPs. This indicates that bipartite binding is widespread, the two motif cores are often identical, and the bound sequences are low in complexity. **(3)** We made BMF available both as a command-line tool with detailed documentation (see A1.4) and as a web server. The server minimizes the technical skills required for analysing new interaction datasets for further discovery

of bipartite binding behavior. **(4)** We introduce the first available method to predict the total K_d of RBPs with any number of binding sites based on K_d values of its individual domains. Using this model we show how cooperativity achieves highly specific and affine interactions, and how small changes in motif density can regulate the affinity of RBPs.

5.3 Future challenges

Characterization of RNA degradation pathways

In recent years our understanding of how RNA degradation is regulated has considerably improved. However, central aspects of RNA recognition and targeting remains obscure. For instance, many RBPs can associate with the degradation machineries to influence their function. An example is the yeast Mmi1 RBP that associates with the Ccr4/Not complex to suppress meiosis transcripts in the normal vegetative growth phase (Stowell et al., 2016). Similarly, other RBPs may be involved to facilitate/constrain the recruitment of degradation factors to sub-classes of RNA molecules. It is crucial to identify and characterize such binding proteins in order to understand the regulation principles of RNA homeostasis.

Another aspect of RNA recognition and targeting that remain poorly understood is their role of the exonucleases in preprocessing of nuclear ncRNAs. It is not clear how the preprocessing stops at specific nucelotides for some RNAs, whereas others can undergo full RNA degradation. While it has been suggested that structural constraints imposed by the RNP could stop the exosome complex at specific positions (Makino et al., 2015), the principles that regulate this process remain elusive.

In the first part of this thesis, we identified TRAMP4 and exosome as the main players of the Nrd1/Nab3 pathway for RNA surveillance in yeast. The TRAMP complex in yeast has many similarities with the nuclear exosome targeting (NEXT) complex in humans (Lubas et al., 2015). Experimental depletion of NEXT subunits was shown to stabilize promoter upstream transcripts and create unprocessed snRNAs (Ntini et al., 2013; Shcherbik et al., 2010). However, how the NEXT complex targets the aberrant human Pol II transcripts is not clear as no homologous complex to the yeast Nrd1/Nab3 system has been identified in metazoans. Therefore, characterizing the players of the mammalian transcription surveillance machinery is an ongoing challenge. New RNA-centric approaches for the discovery of RBPs (such as Trendel et al., 2019; Castello et al., 2016) may pave the way for characterization of new mammalian degradation pathways in the future.

Another interesting aspect of RNA degradation is the enrichment of many of its components in the phase separated cytosolic Processing-bodies (P-bodies) upon stress (Zhang and Herman, 2020). P-bodies enrich specific mRNAs and proteins, including Xrn1 and decapping proteins (Eulalio et al., 2007; Youn et al., 2019). Although they are evolutionarily conserved, their biological function is not yet fully understood. The early discovery of RNA processing enzymes in P-granules lead to the believe that these granules are sites of active mRNA decay (Balagopal and Parker, 2009). However, evidence has accumulated in recent years to indicate that mRNA decay may even be suppressed in these granules and some mRNAs can get stored in P-bodies over long times (Standart and Weil, 2018; Hubstenberger et al., 2017; Arribere et al., 2011; Huch and Nissan, 2017). It will be therefore crucial to characterize the activity of P-body proteins and the cellular effect of their localization to shed light on the function of these RNP granules.

Understanding the mRNP code

RBPs are crucial for most biological processes and the complexity of their regulation grows with with organism complexity. Recently, 7.5% of the human proteome was estimated to constitute RBPs (Gerstberger et al., 2014), a similar fraction as DNA-binding factors. Considering that many of these RBPs are as yet uncharacterized and that the regulatory pathways behind many RNA processing systems remain elusive, it seems that our knowledge of RBP functions is only the tip of the iceberg. I will therefore mention a few areas that may benefit from further developments.

Most current computational methods for RNA motif discovery take RNA sequence, secondary structure, and sometimes the regional information of the binding sites as input parameters (Sasse et al., 2018; Yan and Zhu, 2020; Pan et al., 2019). We show in this work that incorporating cooperativity into RNA motif models would further enhance their prediction accuracy. Another improvement could come from the way structure is incorporated into RBP models. The secondary structure is computationally predicted from input sequences (Sasse et al., 2018; Maticzka et al., 2014; Pan et al., 2018). Computational estimation of RNA structure often assumes that the molecule will fold into its minimum free energy state (Miao and Westhof, 2017; Seetin and Mathews, 2012; Laing and Schlick, 2011). The crowded cellular environment, however, challenges this assumption as RNAs are often bound by many RBPs that restrain their folding dynamics (Gehring et al., 2017). The differential expression of bound RBPs in various cell types and environments may consequently alter RNA structure and therefore further complicate the structural prediction. I therefore expect that new structure prediction algorithms that take the RNA and protein concentrations into account (by incorporating proteomic and transcriptomic data) could more accurately reflect the RNA structure in a given cell. Furthermore, experimental approaches to global mapping of RNA structures in living cells, such as SHAPE (Merino et al., 2005; Kertesz et al., 2010; Talkish et al., 2014), could directly provide the data needed to train better RNA motif models, as well as contributing valuable training data for advancing computational structure prediction algorithms. Another unexplored property of RNAs for motif discovery are RNA modifications that could alter RBP binding. For instance adenosine methylation was shown to influence whether or not RNA binding sites are available to hnRNPC (Liu et al., 2015). With the availability of high-throughput techniques for mapping RNA modifications (Helm and Motorin, 2017; Jonkhout et al., 2017), these datasets could contribute another rich resource for training better RBP interaction models.

Biochemical studies could furthermore get deployed to validate and characterize the observations made by our thermodynamic models. These assays could study the effect of RNA linker length, motif density on RNA, and the number of RBDs on the overall K_{d}. A new method was recently developed for measuring K_{d} values in cells using high-throughput sequencing (Sharma et al., 2021). This technique could pave the road for large-scale measurements of RNA-binding dynamics for many RBPs inside living cells and could answer some of the long standing questions in the field, such as how the binding dynamics of an RBP differs between various target transcripts, how long RBPs reside on their targets, and how many different proteins bind individual mRNAs at a given time.

Another exciting advancement for studying RBPs in the last decade is the development of many technologies for large-scale identification of RBPs in cells, many of which are non-canonical RBPs. (Schmidt et al., 2012; Trendel et al., 2019; Castello et al., 2016; Hentze et al., 2018). Studying binding specificities of these non-conventional RBPs and identifying their endogenous targets could expand our understanding of RNA regulation or perhaps provide new insights on how RNAs can in turn regulate their target proteins. Interestingly many of the RBPs identified tend to be enriched in intrinsically disordered regions (Castello et al., 2016; Balcerak et al., 2019). Yet, most structural and biochemical studies limit their scope to characterizing the ordered regions in these proteins. Further investigations are needed to

understand the role of disordered regions in binding and specificity of RBPs.

Finally, most current computational and experimental methods for studying RNA-protein interactions are strongly biased towards the protein-centric notion, in which proteins regulate the activity of transcripts and not the other way around. However, recent characterization of many ncRNAs challenges this viewpoint. For instance several functionally important long ncRNAs, such as Xist, have been shown to scaffold the RNA-binding PRC2 complex that facilitates the assembly of histone modification enzymes leading to epigenetically silence specific sets of genes (Davidovich and Cech, 2015; Balcerak et al., 2019). We therefore have to acknowledge the role of RNA molecules in regulating the activity of their bound protein partners. Characterization of many ncRNA transcripts with unknown functions may broaden our understanding of how RNAs fine-tune transcription and regulate protein activity. For this purpose, RNA-centric computational methods can be developed to predict potential interaction partners and shed light on the function of uncharacteristic ncRNAs.

Decoding the molecular grammar of phase separation

The function of many RBPs is controlled through their enrichment and localization in ribonucleic-protein condensates. Multivalent interactions between condensate components are known to promote their formation (Wang et al., 2018; Burke et al., 2015; Dignon et al., 2020; Krainer et al., 2021). However, the forces that uniquely characterize each condensate are unknown. A multitude of biomolecular condensate types have been identified in the cytoplasm and in the nucleus, and many form simultaneously and/or have sub-compartments that do not mix. This suggests the existence of distinct chemical features that give each condensate its unique signature and leads to the "sorting" of each macromolecule to its correct target condensate. Future experimental efforts to generate quantitative maps of condensate constituents could go hand in hand with computational techniques to find the phase separation grammar encoded in protein sequences. In line with this, many computational techniques have been developed to predict the phase separation behavior of individual proteins (Vernon and Forman-Kay, 2019; Vernon et al., 2018; van Mierlo et al., 2021; Hardenberg et al., 2020; Sun et al., 2019; Orlando et al., 2019). While these methods can be valuable for predicting the potential for phase separation, cellular condensation relies on many factors such as the availability of RNA molecules and the concentration of other proteins. Therefore the validity of LLPS-negative datasets that are used by these methods is debatable. With the move towards better characterization of condensate components (via mass spectroscopy of purified granules: Jain et al., image-based screens: Berchtold et al., APEX-MS: Markmiller et al., and APEX-seq: Padron et al.), new computational methods could be developed that take the concentration of RNAs and proteins into account to better predict which macromolecules would get enriched in which biomolecular condensate types.

References

Alipanahi, B., Delong, A., Weirauch, M. T., and Frey, B. J. (2015). Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. *Nature Biotechnology*, 33(8):831–838.

Allmang, C., Kufel, J., Chanfreau, G., Mitchell, P., Petfalski, E., and Tollervey, D. (1999a). Functions of the exosome in rRNA, snoRNA and snRNA synthesis. *The EMBO Journal*, 18(19):5399–5410.

Allmang, C., Petfalski, E., Podtelejnikov, A., Mann, M., Tollervey, D., and Mitchell, P. (1999b). The yeast exosome and human PM–Scl are related complexes of 3' to 5' exonucleases. *Genes & Development*, 13(16):2148–2158.

Anderson, J. T. and Wang, X. (2009). Nuclear RNA surveillance: no sign of substrates tailing off. *Critical reviews in biochemistry and molecular biology*, 44(1):16–24.

Änkö, M.-L. and Neugebauer, K. M. (2012). RNA-protein interactions in vivo: global gets specific. *Trends Biochem. Sci*, 37(7):255–262.

Arigo, J. T., Eyler, D. E., Carroll, K. L., and Corden, J. L. (2006). Termination of Cryptic Unstable Transcripts Is Directed by Yeast RNA-Binding Proteins Nrd1 and Nab3. *Molecular Cell*, 23(6):841–851.

Arribere, J. A., Doudna, J. A., and Gilbert, W. V. (2011). Reconsidering movement of eukaryotic mRNAs between polysomes and P bodies. *Molecular Cell*, 44(5):745–758.

Avsec, Z., Barekatain, M., Cheng, J., and Gagneur, J. (2017). Modeling positional effects of regulatory sequences with spline transformations increases prediction accuracy of deep neural networks. *Bioinformatics*, 34(8):1261–1269.

Azzouz, N., Panasenko, O. O., Deluen, C., Hsieh, J., Theiler, G., and Collart, M. A. (2009). Specific roles for the Ccr4-Not complex subunits in expression of the genome. *RNA*, 15(3):377–383.

Babitzke, P., Baker, C. S., and Romeo, T. (2009). Regulation of Translation Initiation by RNA Binding Proteins. *Annual Review of Microbiology*, 63(1):27–44. PMID: 19385727.

Bachellerie, J.-P., Cavaillé, J., and Hüttenhofer, A. (2002). The expanding snoRNA world. *Biochimie*, 84(8):775–790.

Baejen, C., Andreani, J., Torkler, P., Battaglia, S., Schwalb, B., Lidschreiber, M., Maier, K. C., Boltendahl, A., Rus, P., Esslinger, S., Söding, J., and Cramer, P. (2017). "Genome-wide Analysis of RNA Polymerase II Termination at Protein-Coding Genes". *Molecular Cell*, 66(1):38 – 49.e6.

Baejen, C., Torkler, P., Gressel, S., Essig, K., Söding, J., and Cramer, P. (2014). Transcriptome Maps of mRNP Biogenesis Factors Define Pre-mRNA Recognition. *Molecular Cell*, 55(5):745–757.

Bahrami-Samani, E., Penalva, L. O., Smith, A. D., and Uren, P. J. (2014). Leveraging cross-link modification events in CLIP-seq for motif discovery. *Nucleic Acids Research*, 43(1):95–103.

Bailey, T. L., Johnson, J., Grant, C. E., and Noble, W. S. (2015). The MEME Suite. *Nucleic Acids Research*, 43(W1):W39–W49.

Balagopal, V. and Parker, R. (2009). Polysomes, P bodies and stress granules: states and fates of eukaryotic mRNAs. *Current Opinion in Cell Biology*, 21(3):403–408.

Balcerak, A., Trebinska-Stryjewska, A., Konopinski, R., Wakula, M., and Grzybowska, E. A. (2019). RNA–protein interactions: disorder, moonlighting and junk contribute to eukaryotic complexity. *Open biology*, 9(6):190096.

Banani, S. F., Lee, H. O., Hyman, A. A., and Rosen, M. K. (2017). Biomolecular condensates: organizers of cellular biochemistry. *Nature Reviews Molecular Cell Biology*, 18(5):285–298.

Banjade, S. and Rosen, M. K. (2014). Phase transitions of multivalent proteins can promote clustering of membrane receptors. *eLife*, 3:e04123.

Bartel, D. P. (2018). Metazoan MicroRNAs. *Cell*, 173(1):20–51.

Battaglia, S., Lidschreiber, M., Baejen, C., Torkler, P., Vos, S. M., and Cramer, P. (2017). RNA-dependent chromatin association of transcription elongation factors and Pol II CTD kinases. *eLife*, 6:e25637.

Bazzini, A. A., del Viso, F., Moreno-Mateos, M. A., Johnstone, T. G., Vejnar, C. E., Qin, Y., Yao, J., Khokha, M. K., and Giraldez, A. J. (2016). Codon identity regulates mRNA stability and translation efficiency during the maternal-to-zygotic transition. *The EMBO Journal*, 35(19):2087–2103.

Ben-Bassat, I., Chor, B., and Orenstein, Y. (2018). A deep neural network approach for learning intrinsic protein-RNA binding preferences. *Bioinformatics*, 34(17):i638–i646.

Bentley, D. L. (2014). Coupling mRNA processing with transcription in time and space. *Nature Reviews Genetics*, 15(3):163–175.

Berchtold, D., Battich, N., and Pelkmans, L. (2018). A Systems-Level Study Reveals Regulators of Membrane-less Organelles in Human Cells. *Molecular Cell*, 72(6):1035–1049.e5.

Berretta, J., Pinskaya, M., and Morillon, A. (2008). A cryptic unstable transcript mediates transcriptional trans-silencing of the Ty1 retrotransposon in S. cerevisiae. *Genes & Development*, 22(5):615–626.

Berry, J., Weber, S. C., Vaidya, N., Haataja, M., and Brangwynne, C. P. (2015). RNA transcription modulates phase transition-driven nuclear body assembly. *Proceedings of the National Academy of Sciences*, 112(38):E5237–E5245.

Boehning, M., Dugast-Darzacq, C., Rankovic, M., Hansen, A. S., Yu, T., Marie-Nelly, H., McSwiggen, D. T., Kokic, G., Dailey, G. M., Cramer, P., Darzacq, X., and Zweckstetter, M. (2018). RNA polymerase II clustering through carboxy-terminal domain phase separation. *Nature Struct. Mol. Biol.*, 25(9):833.

Boeynaems, S., Alberti, S., Fawzi, N. L., Mittag, T., Polymenidou, M., Rousseau, F., Schymkowitz, J., Shorter, J., Wolozin, B., Van Den Bosch, L., Tompa, P., and Fuxreiter, M. (2018). Protein phase separation: A new phase in cell biology. *Trends in Cell Biology*, 28(6):420 – 435.

Boija, A., Klein, I. A., Sabari, B. R., Dall'Agnese, A., Coffey, E. L., Zamudio, A. V., Li, C. H., Shrinivas, K., Manteiga, J. C., Hannett, N. M., et al. (2018). Transcription factors activate genes through the phase-separation capacity of their activation domains. *Cell*, 175(7):1842–1855.

Boo, S. H. and Kim, Y. K. (2020). The emerging role of RNA modifications in the regulation of mRNA stability. *Experimental & Molecular Medicine*, 52(3):400–408.

Brangwynne, C. P., Eckmann, C. R., Courson, D. S., Rybarska, A., Hoege, C., Gharakhani, J., Jülicher, F., and Hyman, A. A. (2009). Germline P Granules Are Liquid Droplets That Localize by Controlled Dissolution/Condensation. *Science*, 324(5935):1729–1732.

Brangwynne, C. P., Mitchison, T. J., and Hyman, A. A. (2011). Active liquid-like behavior of nucleoli determines their size and shape in Xenopus laevis oocytes. *Proceedings of the National Academy of Sciences*, 108(11):4334–4339.

Brangwynne, C. P., Tompa, P., and Pappu, R. V. (2015). Polymer physics of intracellular phase transitions. *Nature Physics*, 11(11):899–904.

Brannan, K. W. and Yeo, G. W. (2016). From Protein-RNA Predictions toward a Peptide-RNA Code. *Molecular Cell*, 64(3):437–438.

Braun, J. E., Truffault, V., Boland, A., Huntzinger, E., Chang, C.-T., Haas, G., Weichenrieder, O., Coles, M., and Izaurralde, E. (2012). A direct interaction between DCP1 and XRN1 couples mRNA decapping to 5 exonucleolytic degradation. *Nature Structural & Molecular Biology*, 19(12):1324.

Broach, J. R. (2012). Nutritional control of growth and development in yeast. *Genetics*, 192(1):73–105.

Brown, C. E. and Sachs, A. B. (1998). Poly(A) Tail Length Control in Saccharomyces cerevisiae Occurs by Message-Specific Deadenylation. *Molecular and Cellular Biology*, 18(11):6548–6559.

Brown, J. T., Bai, X., and Johnson, A. W. (2000). The yeast antiviral proteins Ski2p, Ski3p, and Ski8p exist as a complex in vivo. *RNA*, 6(3):449–457.

Budach, S. and Marsico, A. (2018). pysster: classification of biological sequences by learning sequence and structure motifs with convolutional neural networks. *Bioinformatics*, 34(17):3035–3037.

Burke, K. A., Janke, A. M., Rhine, C. L., and Fawzi, N. L. (2015). Residue-by-residue view of in vitro FUS granules that bind the C-terminal domain of RNA polymerase II. *Molecular cell*, 60(2):231–241.

Calabretta, S. and Richard, S. (2015). Emerging Roles of Disordered Sequences in RNA-Binding Proteins. *Trends in Biochemical Sciences*, 40(11):662–672.

Callahan, K. P. and Butler, J. S. (2010). TRAMP Complex Enhances RNA Degradation by the Nuclear Exosome Component Rrp62. *Journal of Biological Chemistry*, 285(6):3540–3547.

Cantara, W. A., Crain, P. F., Rozenski, J., McCloskey, J. A., Harris, K. A., Zhang, X., Vendeix, F. A. P., Fabris, D., and Agris, P. F. (2010). The RNA modification database, RNAMDB: 2011 update. *Nucleic Acids Research*, 39(suppl_1):D195–D201.

Carroll, J. S., Munchel, S. E., and Weis, K. (2011). The DExD/H box ATPase Dhh1 functions in translational repression, mRNA decay, and processing body dynamics. *Journal of Cell Biology*, 194(4):527–537.

Case, L. B., Ditlev, J. A., and Rosen, M. K. (2019). Regulation of Transmembrane Signaling by Phase Separation. *Ann. Rev. Biophysics*, 48:465–494.

Castello, A., Fischer, B., Frese, C. K., Horos, R., Alleaume, A.-M., Foehr, S., Curk, T., Krijgsveld, J., and Hentze, M. W. (2016). Comprehensive Identification of RNA-Binding Domains in Human Cells. *Molecular Cell*, 63(4):696–710.

Cech, T. R. (2012). The RNA Worlds in Context. *Cold Spring Harbor Perspectives in Biology*, 4(7).

Cheng, H., Dufu, K., Lee, C.-S., Hsu, J. L., Dias, A., and Reed, R. (2006). Human mRNA export machinery recruited to the 5' end of mRNA. *Cell*, 127(7):1389–1400.

Cho, W.-K., Spille, J.-H., Hecht, M., Lee, C., Li, C., Grube, V., and Cisse, I. I. (2018). Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science*, 361(6400):412–415.

Choe, J., Oh, N., Park, S., Lee, Y. K., Song, O.-K., Locker, N., Chi, S.-G., and Kim, Y. K. (2012). Translation initiation on mRNAs bound by nuclear cap-binding protein complex CBP80/20 requires interaction between CBP80/20-dependent translation initiation factor and eukaryotic translation initiation factor 3g. *Journal of Biological Chemistry*, 287(22):18500–18509.

Coller, J. and Parker, R. (2005). General translational repression by activators of mrna decapping. *Cell*, 122(6):875 – 886.

Comoglio, F., Sievers, C., and Paro, R. (2015). Sensitive and highly resolved identification of RNA-protein interaction sites in PAR-CLIP data. *BMC Bioinformatics*, 16(1):1–10.

Conduit, P. T., Feng, Z., Richens, J. H., Baumbach, J., Wainman, A., Bakshi, S. D., Dobbelaere, J., Johnson, S., Lea, S. M., and Raff, J. W. (2014). The centrosome-specific phosphorylation of cnn by Polo/Plk1 drives Cnn scaffold assembly and centrosome maturation. *Developmental Cell*, 28:659–669.

Cook, K. B., Vembu, S., Ha, K. C., Zheng, H., Laverty, K. U., Hughes, T. R., Ray, D., and Morris, Q. D. (2017). RNAcompete-S: Combined RNA sequence/structure preferences for RNA binding proteins derived from a single-step in vitro selection. *Methods*, 126:18–28.

Corcoran, D. L., Georgiev, S., Mukherjee, N., Gottwein, E., Skalsky, R. L., Keene, J. D., and Ohler, U. (2011). PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biology*, 12(8):1–16.

Cramer, P. (2019). Organization and regulation of gene transcription. *Nature*, 573(7772):45–54.

Crick, F. (1958). On protein synthesis. *Symposia of the Society for Experimental Biology*, 12:138–63.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563.

Cusack, S. (1997). Aminoacyl-tRNA synthetases. *Current Opinion in Structural Biology*, 7(6):881–889.

Dahm, R. (2005). Friedrich Miescher and the discovery of DNA. *Developmental Biology*, 278(2):274–288.

Danin-Kreiselman, M., Lee, C. Y., and Chanfreau, G. (2003). RNAse III-Mediated Degradation of Unspliced Pre-mRNAs and Lariat Introns. *Molecular Cell*, 11(5):1279–1289.

Davidovich, C. and Cech, T. R. (2015). The recruitment of chromatin modifiers by long noncoding RNAs: lessons from PRC2. *RNA*, 21(12):2007–2022.

Davis, C. A. and Ares, M. (2006). Accumulation of unstable promoter-associated transcripts upon loss of the nuclear exosome subunit Rrp6p in Saccharomycescerevisiae. *Proceedings of the National Academy of Sciences*, 103(9):3262–3267.

Decatur, W. A. and Fournier, M. J. (2003). RNA-guided nucleotide modification of ribosomal and other RNAs. *Journal of Biological Chemistry*, 278(2):695–698.

Delan-Forino, C., Spanos, C., Rappsilber, J., and Tollervey, D. (2020). Substrate specificity of the TRAMP nuclear surveillance complexes. *Nature Communications*, 11(1):1–15.

Dendooven, T., Luisi, B. F., and Bandyra, K. J. (2020). RNA lifetime control, from stereochemistry to gene expression. *Current opinion in structural biology*, 61:59–70.

Deng, L., Liu, Y., Shi, Y., Zhang, W., Yang, C., and Liu, H. (2020). Deep neural networks for inferring binding sites of RNA-binding proteins by using distributed representations of RNA primary sequence and secondary structure. *BMC genomics*, 21(13):1–10.

Dignon, G. L., Best, R. B., and Mittal, J. (2020). Biomolecular phase separation: From molecular driving forces to macroscopic properties. *Annual review of physical chemistry*, 71:53–75.

Djuranovic, S., Nahvi, A., and Green, R. (2011). A Parsimonious Model for Gene Regulation by miRNAs. *Science*, 331(6017):550–553.

Doma, M. K. and Parker, R. (2007). Rna quality control in eukaryotes. *Cell*, 131(4):660 – 668.

Dominguez, D., Freese, P., Alexis, M. S., Su, A., Hochman, M., Palden, T., Bazile, C., Lambert, N. J., Van Nostrand, E. L., Pratt, G. A., et al. (2018). Sequence, structure, and context preferences of human RNA binding proteins. *Molecular cell*, 70(5):854–867.

Dreyfuss, G., Kim, V. N., and Kataoka, N. (2002). Messenger-RNA-binding proteins and the messages they carry. *Nature Reviews Molecular Cell Biology*, 3(3):195–205.

Dunn, E. F., Hammell, C. M., Hodge, C. A., and Cole, C. N. (2005). Yeast poly(A)-binding protein, Pab1, and PAN, a poly(A) nuclease complex recruited by Pab1, connect mRNA biogenesis to export. *Genes & Development*, 19(1):90–103.

Eddy, S. R. (2001). Non–coding RNA genes and the modern RNA world. *Nature Reviews Genetics*, 2(12):919–929.

Eliscovich, C. and Singer, R. H. (2017). RNP transport in cell biology: the long and winding road. *Current Opinion in Cell Biology*, 45:38 – 46. Cell Regulation.

Elkon, R., Ugalde, A. P., and Agami, R. (2013). Alternative cleavage and polyadenylation: extent, regulation and function. *Nature Reviews Genetics*, 14(7):496–506.

Erijman, A., Kozlowski, L., Sohrabi-Jahromi, S., Fishburn, J., Warfield, L., Schreiber, J., Noble, W. S., Söding, J., and Hahn, S. (2020). A High-Throughput screen for transcription activation domains reveals their sequence features and permits prediction by deep learning. *Molecular Cell*, 78(5):890–902.

Eulalio, A., Behm-Ansmant, I., and Izaurralde, E. (2007). P bodies: at the crossroads of post-transcriptional pathways. *Nature Reviews Molecular Cell Biology*, 8(1):9–22.

Falk, S., Weir, J. R., Hentschel, J., Reichelt, P., Bonneau, F., and Conti, E. (2014). The Molecular Architecture of the TRAMP Complex Reveals the Organization and Interplay of Its Two Catalytic Activities. *Molecular Cell*, 55(6):856–867.

Fasken, M. B., Laribee, R. N., and Corbett, A. H. (2015). Nab3 Facilitates the Function of the TRAMP Complex in RNA Processing via Recruitment of Rrp6 Independent of Nrd1. *PLOS Genetics*, 11(3):1–34.

Feric, M., Vaidya, N., Harmon, T. S., Mitrea, D. M., Zhu, L., Richardson, T. M., Kriwacki, R. W., Pappu, R. V., and Brangwynne, C. P. (2016). Coexisting Liquid Phases Underlie Nucleolar Subcompartments. *Cell*, 165(7):1686–1697.

Flaherty, S. M., Fortes, P., Izaurralde, E., Mattaj, I. W., and Gilmartin, G. M. (1997). Participation of the nuclear cap binding complex in pre-mRNA 3' processing. *Proceedings of the National Academy of Sciences*, 94(22):11893–11898.

Flory, P. J. (1942). Thermodynamics of high polymer solutions. *The Journal of chemical physics*, 10(1):51–61.

Fortes, P., Inada, T., Preiss, T., Hentze, M. W., Mattaj, I. W., and Sachs, A. B. (2000). The yeast nuclear cap binding complex can interact with translation factor eIF4G and mediate translation initiation. *Molecular Cell*, 6(1):191–196.

Fox, A. H., Nakagawa, S., Hirose, T., and Bond, C. S. (2018a). Paraspeckles: Where Long Noncoding RNA Meets Phase Separation. *Trends in Biochemical Sciences*, 43(2):124–135.

Fox, A. H., Nakagawa, S., Hirose, T., and Bond, C. S. (2018b). Paraspeckles: where long noncoding RNA meets phase separation. *Trends in Biochemical Sciences*, 43(2):124–135.

Fox, M. J., Gao, H., Smith-Kinnaman, W. R., Liu, Y., and Mosley, A. L. (2015). The Exosome Component Rrp6 Is Required for RNA Polymerase II Termination at Specific Targets of the Nrd1-Nab3 Pathway. *PLOS Genetics*, 11(2):1–26.

Frey, S., Richter, R. P., and Görlich, D. (2006). FG-Rich Repeats of Nuclear Pore Proteins Form a Three-Dimensional Meshwork with Hydrogel-Like Properties. *Science*, 314(5800):815–817.

Friedersdorf, M. B. and Keene, J. D. (2014). Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs. *Genome Biology*, 15(1):1–16.

Frith, M. C., Saunders, N. F., Kobe, B., and Bailey, T. L. (2008). Discovering sequence motifs with arbitrary insertions and deletions. *PLoS Computational Biology*, 4(5):e1000071.

Gagnon, J. A. and Mowry, K. L. (2011). Molecular motors: directing traffic during RNA localization. *Critical Reviews in Biochemistry and Molecular Biology*, 46(3):229–239.

Galganski, L., Urbanek, M. O., and Krzyzosiak, W. J. (2017). Nuclear speckles: molecular organization, biological function and role in disease. *Nucleic Acids Research*, 45(18):10350–10368.

Garzia, A., Meyer, C., Morozov, P., Sajek, M., and Tuschl, T. (2017). Optimization of PAR-CLIP for transcriptome-wide identification of binding sites of RNA-binding proteins. *Methods*, 118-119:24–40. Protein-RNA: Structure Function and Recognition.

Gehring, N. H., Wahle, E., and Fischer, U. (2017). Deciphering the mRNP Code: RNA-Bound Determinants of Post-Transcriptional Gene Regulation. *Trends in Biochemical Sciences*, 42(5):369–382.

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

Ghanbari, M. and Ohler, U. (2020). Deep neural networks for interpreting RNA-binding protein target preferences. *Genome Research*, 30(2):214–226.

Gilbert, C. and Svejstrup, J. Q. (2006). RNA immunoprecipitation for determining RNA-protein associations in vivo. *Current Protocols in Molecular Biology*, 75(1):27–4.

Gilbert, W. (1986). Origin of life: The RNA world. *nature*, 319(6055):618–618.

Gilbert, W. V., Bell, T. A., and Schaening, C. (2016). Messenger RNA modifications: Form, distribution, and function. *Science*, 352(6292):1408–1412.

Gonatopoulos-Pournatzis, T. and Cowling, V. H. (2014). Cap-binding complex (CBC). *Biochemical Journal*, 457(2):231–242.

Görnemann, J., Kotovic, K. M., Hujer, K., and Neugebauer, K. M. (2005). Cotranscriptional spliceosome assembly occurs in a stepwise fashion and requires the cap binding complex. *Molecular Cell*, 19(1):53–63.

Green, M. R. (1986). Pre-mRNA splicing. *Annual review of genetics*, 20(1):671–708.

Gruber, A. J. and Zavolan, M. (2019). Alternative cleavage and polyadenylation in health and disease. *Nature Reviews Genetics*, 20(10):599–614.

Guo, Y. E., Manteiga, J. C., Henninger, J. E., Sabari, B. R., Dall'Agnese, A., Hannett, N. M., Spille, J.-H., Afeyan, L. K., Zamudio, A. V., Shrinivas, K., et al. (2019). Pol II phosphorylation regulates a switch between transcriptional and splicing condensates. *Nature*, 572(7770):543–548.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano Jr, M., Jungkamp, A.-C., Munschauer, M., et al. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1):129–141.

Hager, G. L., McNally, J. G., and Misteli, T. (2009). Transcription dynamics. *Molecular Cell*, 35(6):741–753.

Hahn, S. (2018). Phase separation, protein disorder, and enhancer function. *Cell*, 175(7):1723–1725.

Hamill, S., Wolin, S. L., and Reinisch, K. M. (2010). Structure and function of the polymerase core of TRAMP, a RNA surveillance complex. *Proceedings of the National Academy of Sciences*, 107(34):15045–15050.

Hanson, G. and Coller, J. (2018). Codon optimality, bias and usage in translation and mRNA decay. *Nature Reviews Molecular Cell Biology*, 19(1):20–30.

Hardenberg, M., Horvath, A., Ambrus, V., Fuxreiter, M., and Vendruscolo, M. (2020). Widespread occurrence of the droplet state of proteins in the human proteome. *Proceedings of the National Academy of Sciences*, 117(52):33254–33262.

Hartmann, H., Guthöhrlein, E. W., Siebert, M., Luehr, S., and Söding, J. (2013). P-value-based regulatory motif discovery using positional weight matrices. *Genome Research*, 23(1):181–194.

Haseloff, J. and Gerlach, W. L. (1988). Simple RNA enzymes with new and highly specific endoribonuclease activities. *Nature*, 334(6183):585–591.

Hashim, F. A., Mabrouk, M. S., and Al-Atabany, W. (2019). Review of different sequence motif finding algorithms. *Avicenna Journal of Medical Biotechnology*, 11(2):130.

He, F., Celik, A., Wu, C., and Jacobson, A. (2018). General decapping activators target different subsets of inefficiently translated mRNAs. *eLife*, 7:e34409.

He, F. and Jacobson, A. (2015). Nonsense-Mediated mRNA Decay: Degradation of Defective Transcripts Is Only Part of the Story. *Annual Review of Genetics*, 49(1):339–366. PMID: 26436458.

He, Y., Fang, J., Taatjes, D. J., and Nogales, E. (2013). Structural visualization of key steps in human transcription initiation. *Nature*, 495(7442):481–486.

Heller, D., Krestel, R., Ohler, U., Vingron, M., and Marsico, A. (2017). ssHMM: extracting intuitive sequence-structure motifs from high-throughput RNA-binding protein data. *Nucleic Acids Research*, 45(19):11004–11018.

Helm, M. and Motorin, Y. (2017). Detecting RNA modifications in the epitranscriptome: predict and validate. *Nature Reviews Genetics*, 18(5):275–291.

Hennig, J. and Sattler, M. (2015). Deciphering the protein-RNA recognition code: Combining large-scale quantitative methods with structural biology. *BioEssays*, 37(8):899–908.

Henras, A. K., Plisson-Chastang, C., O'Donohue, M.-F., Chakraborty, A., and Gleizes, P.-E. (2015). An overview of pre-ribosomal RNA processing in eukaryotes. *Wiley Interdisciplinary Reviews: RNA*, 6(2):225–242.

Hentze, M. W., Castello, A., Schwarzl, T., and Preiss, T. (2018). A brave new world of RNA-binding proteins. *Nature Reviews Molecular Cell Biology*, 19(5):327.

Herzel, L., Ottoz, D. S., Alpert, T., and Neugebauer, K. M. (2017). Splicing and transcription touch base: co-transcriptional spliceosome assembly and function. *Nature Reviews Molecular Cell Biology*, 18(10):637.

Hiller, M., Pudimat, R., Busch, A., and Backofen, R. (2006). Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Research*, 34(17):e117–e117.

Holley, R. W., Apgar, J., Everett, G. A., Madison, J. T., Marquisee, M., Merrill, S. H., Penswick, J. R., and Zamir, A. (1965). Structure of a ribonucleic acid. *Science*, pages 1462–1465.

Hoque, M., Ji, Z., Zheng, D., Luo, W., Li, W., You, B., Park, J. Y., Yehia, G., and Tian, B. (2013). Analysis of alternative cleavage and polyadenylation by 3' region extraction and deep sequencing. *Nature methods*, 10(2):133–139.

Houseley, J., LaCava, J., and Tollervey, D. (2006). RNA-quality control by the exosome. *Nature Reviews Molecular Cell Biology*, 7(7):529–539.

Houseley, J. and Tollervey, D. (2008). The nuclear RNA surveillance machinery: the link between ncRNAs and genome structure in budding yeast? *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1779(4):239–246.

Houseley, J. and Tollervey, D. (2009). The many pathways of RNA degradation. *Cell*, 136(4):763–776.

Hubstenberger, A., Courel, M., Bénard, M., Souquere, S., Ernoult-Lange, M., Chouaib, R., Yi, Z., Morlot, J.-B., Munier, A., Fradet, M., et al. (2017). P-body purification reveals the condensation of repressed mRNA regulons. *Molecular Cell*, 68(1):144–157.

Huch, S. and Nissan, T. (2017). An mRNA decapping mutant deficient in P body assembly limits mRNA stabilization in response to osmotic stress. *Scientific Reports*, 7(1):1–13.

Izaurralde, E., Lewis, J., Gamberi, C., Jarmolowski, A., McGuigan, C., and Mattaj, I. W. (1995). A cap-binding protein complex mediating U snRNA export. *Nature*, 376(6542):709–712.

Jackson, R. J., Hellen, C. U. T., and Pestova, T. V. (2010). The mechanism of eukaryotic translation initiation and principles of its regulation. *Nature Reviews Molecular Cell Biology*, 11(2):113–127.

Jain, S., Wheeler, J. R., Walters, R. W., Agrawal, A., Barsic, A., and Parker, R. (2016). ATPase-modulated stress granules contain a diverse proteome and substructure. *Cell*, 164(3):487–498.

Jankowsky, E. and Harris, M. E. (2015). Specificity and nonspecificity in RNA–protein interactions. *Nature Reviews Molecular Cell Biology*, 16(9):533–544.

Jha, A., Aicher, J. K., Gazzara, M. R., Singh, D., and Barash, Y. (2020). Enhanced Integrated Gradients: improving interpretability of deep learning models using splicing codes as a case study. *Genome Biology*, 21(1):1–22.

Jia, H., Wang, X., Liu, F., Guenther, U.-P., Srinivasan, S., Anderson, J. T., and Jankowsky, E. (2011). The RNA helicase Mtr4p modulates polyadenylation in the TRAMP complex. *Cell*, 145(6):890–901.

Jiao, X., Xiang, S., Oh, C., Martin, C. E., Tong, L., and Kiledjian, M. (2010). Identification of a quality-control mechanism for mRNA 5'-end capping. *Nature*, 467(7315):608–611.

Jinek, M., Coyle, S. M., and Doudna, J. A. (2011). Coupled 5' nucleotide recognition and processivity in Xrn1-mediated mRNA decay. *Molecular Cell*, 41(5):600–608.

Jolma, A., Zhang, J., Mondragón, E., Morgunova, E., Kivioja, T., Laverty, K. U., Yin, Y., Zhu, F., Bourenkov, G., Morris, Q., et al. (2020). Binding specificities of human RNA-binding proteins toward structured and linear RNA sequences. *Genome research*, 30(7):962–973.

Jonkhout, N., Tran, J., Smith, M. A., Schonrock, N., Mattick, J. S., and Novoa, E. M. (2017). The RNA modification landscape in human disease. *RNA*, 23(12):1754–1769.

Julian König (2021). `https://www.bmls.de/Computational_RNA_Biology/aboutus.html`. [Online; accessed 01-February-2021].

Kapranov, P., Cheng, J., Dike, S., Nix, D. A., Duttagupta, R., Willingham, A. T., Stadler, P. F., Hertel, J., Hackermüller, J., Hofacker, I. L., Bell, I., Cheung, E., Drenkow, J., Dumais, E., Patel, S., Helt, G., Ganesh, M., Ghosh, S., Piccolboni, A., Sementchenko, V., Tammana, H., and Gingeras, T. R. (2007). RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription. *Science*, 316(5830):1484–1488.

Kertesz, M., Wan, Y., Mazor, E., Rinn, J. L., Nutter, R. C., Chang, H. Y., and Segal, E. (2010). Genome-wide measurement of RNA secondary structure in yeast. *Nature*, 467(7311):103–107.

Kiesel, A., Roth, C., Ge, W., Wess, M., Meier, M., and Söding, J. (2018). The BaMM web server for de-novo motif discovery and regulatory sequence analysis. *Nucleic Acids Research*, 46(W1):W215–W220.

Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M., and Zavolan, M. (2011). A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nature Methods*, 8(7):559–564.

Kiss, T. (2001). Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. *The EMBO journal*, 20(14):3617–3622.

Kiss, T. (2004). Biogenesis of small nuclear RNPs. *Journal of Cell Science*, 117(25):5949–5951.

Kiss-László, Z., Henry, Y., and Kiss, T. (1998). Sequence and structural elements of methylation guide snoRNAs essential for site-specific ribose methylation of pre-rRNA. *The EMBO Journal*, 17(3):797–807.

König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M., and Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature Struct Mol Biol*, 17(7):909.

Krainer, G., Welsh, T. J., Joseph, J. A., St George-Hyslop, P., Hyman, A. A., Collepardo-Guevara, R., Alberti, S., and Knowles, T. P. (2021). Reentrant Liquid Condensate Phase of Proteins is Stabilized by Hydrophobic and Non-Ionic interactions. *Biophysical Journal*, 120(3):28a.

Kruk, J. A., Dutta, A., Fu, J., Gilmour, D. S., and Reese, J. C. (2011). The multifunctional Ccr4–Not complex directly promotes transcription elongation. *Genes & development*, 25(6):581–593.

LaCava, J., Houseley, J., Saveanu, C., Petfalski, E., Thompson, E., Jacquier, A., and Tollervey, D. (2005). RNA degradation by the exosome is promoted by a nuclear polyadenylation complex. *Cell*, 121(5):713–724.

Lackner, D. H., Schmidt, M. W., Wu, S., Wolf, D. A., and Bähler, J. (2012). Regulation of transcriptome, translation, and proteome in response to environmental stress in fission yeast. *Genome Biology*, 13(4):1–14.

Laing, C. and Schlick, T. (2011). Computational approaches to RNA structure prediction, analysis, and design. *Current Opinion in Structural Biology*, 21(3):306–318.

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T. R., and Weirauch, M. T. (2018). The human transcription factors. *Cell*, 172(4):650–665.

Langdon, E. M. and Gladfelter, A. S. (2018). A New Lens for RNA Localization: Liquid-Liquid Phase Separation. *Annual Review of Microbiology*, 72(1):255–271. PMID: 30200855.

Lardelli, R. M. and Lykke-Andersen, J. (2020). Competition between maturation and degradation drives human snRNA 3' end quality control. *Genes & Development*, 34(13-14):989–1001.

Laribee, R. N., Hosni-Ahmed, A., Workman, J. J., and Chen, H. (2015). Ccr4-Not Regulates RNA Polymerase I Transcription and Couples Nutrient Signaling to the Control of Ribosomal RNA Biogenesis. *PLOS Genetics*, 11(3):1–24.

Levine, M. and Tjian, R. (2003). Transcription regulation and animal diversity. *Nature*, 424(6945):147–151.

Li, X., Kazan, H., Lipshitz, H. D., and Morris, Q. D. (2014). Finding the target sites of RNA-binding proteins. *Wiley Interdisciplinary Reviews: RNA*, 5(1):111–130.

Lianoglou, S., Garg, V., Yang, J. L., Leslie, C. S., and Mayr, C. (2013). Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes & development*, 27(21):2380–2396.

Licatalosi, D. D. and Darnell, R. B. (2010). RNA processing and its regulation: global insights into biological networks. *Nature Reviews Genetics*, 11(1):75–87.

Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W., Clark, T. A., Schweitzer, A. C., Blume, J. E., Wang, X., et al. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221):464–469.

Lincoln, T. A. and Joyce, G. F. (2009). Self-Sustained Replication of an RNA Enzyme. *Science*, 323(5918):1229–1232.

Liu, N., Dai, Q., Zheng, G., He, C., Parisien, M., and Pan, T. (2015). N 6-methyladenosine-dependent RNA structural switches regulate RNA–protein interactions. *Nature*, 518(7540):560–564.

Liu, Q., Greimann, J. C., and Lima, C. D. (2006). Reconstitution, activities, and structure of the eukaryotic RNA exosome. *Cell*, 127(6):1223–1237.

Lodish, H., Berk, A., Zipursky, S. L., Matsudaira, P., Baltimore, D., and Darnell, J. (2000). The three roles of RNA in protein synthesis. In *Molecular Cell Biology. 4th edition*. WH Freeman.

Lubas, M., Andersen, P. ., Schein, A., Dziembowski, A., Kudla, G., and Jensen, T. . (2015). The Human Nuclear Exosome Targeting Complex Is Loaded onto Newly Synthesized RNA to Direct Early Ribonucleolysis. *Cell Reports*, 10(2):178–192.

Lunde, B. M., Moore, C., and Varani, G. (2007). RNA-binding proteins: modular design for efficient function. *Nature Reviews Molecular Cell Biology*, 8(6):479–490.

Madhani, H. D. (2013). snRNA Catalysts in the Spliceosome's Ancient Core. *Cell*, 155(6):1213–1215.

Makino, D. L., Schuch, B., Stegmann, E., Baumgärtner, M., Basquin, C., and Conti, E. (2015). RNA degradation paths in a 12-subunit nuclear exosome complex. *Nature*, 524(7563):54–58.

Markmiller, S., Soltanieh, S., Server, K. L., Mak, R., Jin, W., Fang, M. Y., Luo, E.-C., Krach, F., Yang, D., Sen, A., Fulzele, A., Wozniak, J. M., Gonzalez, D. J., Kankel, M. W., Gao, F.-B., Bennett, E. J., LÃ©cuyer, E., and Yeo, G. W. (2018). Context-Dependent and Disease-Specific Diversity in Protein Interactions within Stress Granules. *Cell*, 172(3):590–604.e13.

Marquardt, S., Hazelbaker, D. Z., and Buratowski, S. (2011). Distinct RNA degradation pathways and 3' extensions of yeast non-coding RNA species. *Transcription*, 2(3):145–154. PMID: 21826286.

Martens, J. A., Laprade, L., and Winston, F. (2004). Intergenic transcription is required to repress the Saccharomyces cerevisiae SER3 gene. *Nature*, 429(6991):571–574.

Martinez-Rucobo, F. W., Kohler, R., van de Waterbeemd, M., Heck, A. J., Hemann, M., Herzog, F., Stark, H., and Cramer, P. (2015). Molecular basis of transcription-coupled pre-mRNA capping. *Molecular Cell*, 58(6):1079–1089.

Matera, A. G., Terns, R. M., and Terns, M. P. (2007). Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nature Reviews Molecular Cell Biology*, 8(3):209–220.

Matera, A. G. and Wang, Z. (2014). A day in the life of the spliceosome. *Nature Reviews Molecular Cell Biology*, 15(2):108–121.

Maticzka, D., Lange, S. J., Costa, F., and Backofen, R. (2014). GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biology*, 15(1):1–18.

Matlin, A. J., Clark, F., and Smith, C. W. (2005). Understanding alternative splicing: towards a cellular code. *Nature Reviews Molecular Cell Biology*, 6(5):386–398.

Mattick, J. S. and Makunin, I. V. (2006). Non-coding RNA. *Human Molecular Genetics*, 15(suppl_1):R17–R29.

Merino, E. J., Wilkinson, K. A., Coughlan, J. L., and 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.

Miao, Z. and Westhof, E. (2017). RNA Structure: Advances and Assessment of 3D Structure Prediction. *Annual Review of Biophysics*, 46(1):483–503. PMID: 28375730.

Miller, C., Schwalb, B., Maier, K., Schulz, D., Dümcke, S., Zacher, B., Mayer, A., Sydow, J., Marcinowski, L., Dölken, L., Martin, D. E., Tresch, A., and Cramer, P. (2011). Dynamic transcriptome analysis measures rates of mRNA synthesis and decay in yeast. *Molecular Systems Biology*, 7(1):458.

Miller, J. E. and Reese, J. C. (2012). Ccr4-Not complex: the control freak of eukaryotic cells. *Critical Reviews in Biochemistry and Molecular Biology*, 47(4):315–333.

Milovanovic, D., Wu, Y., Bian, X., and De Camilli, P. (2018). A liquid phase of synapsin and lipid vesicles. *Science*, 361(6402):604–607.

Mitchell, S. F. and Parker, R. (2014). Principles and properties of eukaryotic mRNPs. *Molecular cell*, 54(4):547–558.

Mittag, T. and Parker, R. (2018). Multiple Modes of Protein-Protein Interactions Promote RNP Granule Assembly. *Journal of Molecular Biology*, 430(23):4636–4649. Phase Separation in Biology and Disease.

Molliex, A., Temirov, J., Lee, J., Coughlin, M., Kanagaraj, A. P., Kim, H. J., Mittag, T., and Taylor, J. P. (2015). Phase Separation by Low Complexity Domains Promotes Stress Granule Assembly and Drives Pathological Fibrillization. *Cell*, 163(1):123–133.

Moss, T., Langlois, F., Gagnon-Kugler, T., and Stefanovsky, V. (2007). A housekeeper with power of attorney: the rRNA genes in ribosome biogenesis. *Cellular and Molecular Life Sciences*, 64(1):29–49.

Muckenthaler, M., Gray, N. K., and Hentze, M. W. (1998). Irp-1 binding to ferritin mrna prevents the recruitment of the small ribosomal subunit by the cap-binding complex eif4f. *Molecular Cell*, 2(3):383 – 388.

Müller-McNicoll, M. and Neugebauer, K. M. (2013). How cells get the message: dynamic assembly and function of mRNA–protein complexes. *Nature Reviews Genetics*, 14(4):275–287.

Munteanu, A., Mukherjee, N., and Ohler, U. (2018). SSMART: sequence-structure motif identification for RNA-binding proteins. *Bioinformatics*, 34(23):3990–3998.

Nam, J.-W., Choi, S.-W., and You, B.-H. (2016). Incredible RNA: dual functions of coding and noncoding. *Molecules and Cells*, 39(5):367.

Neil, H., Malabat, C., d'Aubenton Carafa, Y., Xu, Z., Steinmetz, L. M., and Jacquier, A. (2009). Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature*, 457(7232):1038–1042.

Neuvéglise, C., Marck, C., and Gaillardin, C. (2011). The intronome of budding yeasts. *Comptes rendus biologies*, 334(8-9):662–670.

Nicastro, G., Candel, A. M., Uhl, M., Oregioni, A., Hollingworth, D., Backofen, R., Martin, S. R., and Ramos, A. (2017). Mechanism of β-actin mRNA Recognition by ZBP1. *Cell Reports*, 18(5):1187–1199.

Ntini, E., Järvelin, A. I., Bornholdt, J., Chen, Y., Boyd, M., Jørgensen, M., Andersson, R., Hoof, I., Schein, A., Andersen, P. R., et al. (2013). Polyadenylation site–induced decay of upstream transcripts enforces promoter directionality. *Nature Structural & Molecular Biology*, 20(8):923.

Oberstrass, F. C., Auweter, S. D., Erat, M., Hargous, Y., Henning, A., Wenter, P., Reymond, L., Amir-Ahmady, B., Pitsch, S., Black, D. L., and Allain, F. H.-T. (2005). Structure of PTB Bound to RNA: Specific Binding and Implications for Splicing Regulation. *Science*, 309(5743):2054–2057.

Orenstein, Y. and Shamir, R. (2014). A comparative analysis of transcription factor binding models learned from PBM, HT-SELEX and ChIP data. *Nucleic Acids Research*, 42(8):e63–e63.

Orenstein, Y., Wang, Y., and Berger, B. (2016). RCK: accurate and efficient inference of sequence- and structure-based protein-RNA binding models from RNAcompete data. *Bioinformatics*, 32(12):i351–i359.

Orlando, G., Raimondi, D., Tabaro, F., Codicè, F., Moreau, Y., and Vranken, W. F. (2019). Computational identification of prion-like RNA-binding proteins that form liquid phase-separated condensates. *Bioinformatics*, 35(22):4617–4623.

Ozdilek, B. A., Thompson, V. F., Ahmed, N. S., White, C. I., Batey, R. T., and Schwartz, J. C. (2017). Intrinsically disordered RGG/RG domains mediate degenerate specificity in RNA binding. *Nucleic Acids Research*, 45(13):7984–7996.

Padron, A., Iwasaki, S., and Ingolia, N. T. (2019). Proximity RNA Labeling by APEX-Seq Reveals the Organization of Translation Initiation Complexes and Repressive RNA Granules. *Molecular Cell*, 75(4):875–887.e5.

Pak, C. W., Kosno, M., Holehouse, A. S., Padrick, S. B., Mittal, A., Ali, R., Yunus, A. A., Liu, D. R., Pappu, R. V., and Rosen, M. K. (2016). Sequence determinants of intracellular phase separation by complex coacervation of a disordered protein. *Molecular Cell*, 63(1):72–85.

Pan, X., Rijnbeek, P., Yan, J., and Shen, H.-B. (2018). Prediction of RNA-protein sequence and structure binding preferences using deep convolutional and recurrent neural networks. *BMC genomics*, 19(1):1–11.

Pan, X. and Shen, H.-B. (2017). RNA-protein binding motifs mining with a new hybrid deep learning based cross-domain knowledge integration approach. *BMC bioinformatics*, 18(1):1–14.

Pan, X. and Shen, H.-B. (2018). Predicting RNA–protein binding sites and motifs through combining local and global deep convolutional neural networks. *Bioinformatics*, 34(20):3427–3436.

Pan, X., Yang, Y., Xia, C.-Q., Mirza, A. H., and Shen, H.-B. (2019). Recent methodology progress of deep learning for RNA-protein interaction prediction. *WIREs RNA*, 10(6):e1544.

Park, J., Kang, M., and Kim, M. (2015). Unraveling the mechanistic features of RNA polymerase II termination by the 5'-3' exoribonuclease Rat1. *Nucleic Acids Research*, 43(5):2625–2637.

Parker, R. (2012). RNA Degradation in Saccharomyces cerevisae. *Genetics*, 191(3):671–702.

Pederson, T. (2011). The Nucleolus. *Cold Spring Harbor Perspectives in Biology*, 3(3).

Pemberton, L. F. and Paschal, B. M. (2005). Mechanisms of Receptor-Mediated Nuclear Import and Nuclear Export. *Traffic*, 6(3):187–198.

Porrua, O. and Libri, D. (2015). Transcription termination and the control of the transcriptome: why, where and how to stop. *Nature Reviews Molecular Cell Biology*, 16(3):190–202.

Proudfoot, N. J. (2011). Ending the message: poly (A) signals then and now. *Genes & development*, 25(17):1770–1782.

Ptashne, M. and Gann, A. (1997). Transcriptional activation by recruitment. *Nature*, 386(6625):569–577.

Radhakrishnan, A., Chen, Y.-H., Martin, S., Alhusaini, N., Green, R., and Coller, J. (2016). The DEAD-box protein Dhh1p couples mRNA decay and translation by monitoring codon optimality. *Cell*, 167(1):122–132.

RajBhandary, U. L. (1968). Studies on Polynucleotides: LXXVII. The Labeling of End Groups in Polynucleotide Chains: The Selective Modification of Diol End Groups in Ribonucleic Acids. *Journal of Biological Chemistry*, 243(3):556–564.

Ramakrishnan, V. (2002). Ribosome structure and the mechanism of translation. *Cell*, 108(4):557–572.

Ramanathan, A., Robb, G. B., and Chan, S.-H. (2016). mRNA capping: biological functions and applications. *Nucleic Acids Research*, 44(16):7511–7526.

Ray, D., Kazan, H., Cook, K. B., Weirauch, M. T., Najafabadi, H. S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., et al. (2013). A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–177.

Reese, J. C. (2013). The control of elongation by the yeast Ccr4 Not complex. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1829(1):127–133. RNA polymerase II Transcript Elongation.

Rich, A. and Davies, D. R. (1956). A new two stranded helical structure: polyadenylic acid and polyuridylic acid. *Journal of the American Chemical Society*, 78(14):3548–3549.

Roeder, R. G. (2019). 50+ years of eukaryotic transcription: an expanding universe of factors and mechanisms. *Nature Structural & Molecular Biology*, 26(9):783–791.

Ross, J. (1996). Control of messenger RNA stability in higher eukaryotes. *Trends in Genetics*, 12(5):171 – 175.

Roth, C. and Torkler, P. (2019). Mockinbird pipeline for PAR-CLIP data analysis. `https://github.com/soedinglab/mockinbird`.

Sabari, B. R., Dall'Agnese, A., Boija, A., Klein, I. A., Coffey, E. L., Shrinivas, K., Abraham, B. J., Hannett, N. M., Zamudio, A. V., Manteiga, J. C., Li, C. H., Guo, Y. E., Day, D. S., Schuijers, J., Vasile, E., Malik, S., Hnisz, D., Lee, T. I., Cisse, I. I., Roeder, R. G., Sharp, P. A., Chakraborty, A. K., and Young, R. A. (2018). Coactivator condensation at super-enhancers links phase separation and gene control. *Science*, 361(6400):eaar3958.

Sabari, B. R., Dall'Agnese, A., and Young, R. A. (2020). Biomolecular condensates in the nucleus. *Trends in biochemical sciences*, 45.

Sakharkar, M. K., Perumal, B. S., Sakharkar, K. R., and Kangueane, P. (2005). An analysis on gene architecture in human and mouse genomes. *In Silico Biology*, 5(4):347–365.

Sasse, A., Laverty, K. U., Hughes, T. R., and Morris, Q. D. (2018). Motif models for RNA-binding proteins. *Current Opinion in Structural Biology*, 53:115–123.

Sawyer, I. A., Hager, G. L., and Dundr, M. (2017). Specific genomic cues regulate Cajal body assembly. *RNA Biology*, 14(6):791–803. PMID: 27715441.

Schmid, M. and Jensen, T. H. (2018). Controlling nuclear RNA levels. *Nature Reviews Genetics*, 19(8):518–529.

Schmid, M. and Jensen, T. H. (2019). *The Nuclear RNA Exosome and Its Cofactors*, pages 113–132. Springer International Publishing, Cham.

Schmidt, C., Kramer, K., and Urlaub, H. (2012). Investigation of protein–RNA interactions by mass spectrometry—Techniques and applications. *Journal of proteomics*, 75(12):3478–3494.

Schmidt, H. B. and Görlich, D. (2015). Nup98 FG domains from diverse species spontaneously phase-separate into particles with nuclear pore-like permselectivity. *eLife*, 4:e04251.

Schmidt, K. and Butler, J. S. (2013a). Nuclear RNA surveillance: role of TRAMP in controlling exosome specificity. *WIREs RNA*, 4(2):217–231.

Schmidt, K. and Butler, J. S. (2013b). Nuclear RNA surveillance: role of TRAMP in controlling exosome specificity. *Wiley Interdisciplinary Reviews: RNA*, 4(2):217–231.

Schneider, T., Hung, L.-H., Aziz, M., Wilmen, A., Thaum, S., Wagner, J., Janowski, R., Müller, S., Schreiner, S., Friedhoff, P., et al. (2019). Combinatorial recognition of clustered RNA elements by the multidomain RNA-binding protein IMP3. *Nature communications*, 10(1):1–18.

Schulz, D., Schwalb, B., Kiesel, A., Baejen, C., Torkler, P., Gagneur, J., Soeding, J., and Cramer, P. (2013). Transcriptome Surveillance by Selective Termination of Noncoding RNA Synthesis. *Cell*, 155(5):1075–1087.

Seetin, M. G. and Mathews, D. H. (2012). *RNA Structure Prediction: An Overview of Methods*, pages 99–122. Humana Press, Totowa, NJ.

Seydoux, G. (2018). The p granules of c. elegans: A genetic model for the study of rna-protein condensates. *Journal of Molecular Biology*, 430(23):4702 – 4710. Phase Separation in Biology and Disease.

Sharma, D., Zagore, L. L., Brister, M. M., Ye, X., Crespo-Hernández, C. E., Licatalosi, D. D., and Jankowsky, E. (2021). The kinetic landscape of an RNA-binding protein in cells. *Nature*, pages 1–5.

Shcherbik, N., Wang, M., Lapik, Y. R., Srivastava, L., and Pestov, D. G. (2010). Polyadenylation and degradation of incomplete RNA polymerase I transcripts in mammalian cells. *EMBO reports*, 11(2):106–111.

Shi, H., Wei, J., and He, C. (2019). Where, when, and how: Context-dependent functions of rna methylation writers, readers, and erasers. *Molecular Cell*, 74(4):640 – 650.

Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. (2016). Not just a black box: Learning important features through propagating activation differences. *arXiv*.

Shyu, A.-B., Wilkinson, M. F., and van Hoof, A. (2008). Messenger RNA regulation: to translate or to degrade. *The EMBO Journal*, 27(3):471–481.

Siebert, M. and Soeding, J. (2016). Bayesian Markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Research*, 44(13):6055–6069.

Singh, G., Pratt, G., Yeo, G. W., and Moore, M. J. (2015). The Clothes Make the mRNA: Past and Present Trends in mRNP Fashion. *Annual Review of Biochemistry*, 84(1):325–354. PMID: 25784054.

Sloan, K. E., Schneider, C., and Watkins, N. J. (2012). Comparison of the yeast and human nuclear exosome complexes. *Biochemical Society Transactions*, 40(4):850–855.

Smith, J., Calidas, D., Schmidt, H., Lu, T., Rasoloson, D., and Seydoux, G. (2016). Spatial patterning of P granules by RNA-induced phase separation of the intrinsically-disordered protein MEG-3. *eLife*, 5:e21337.

Sohrabi-Jahromi, S., Hofmann, K. B., Boltendahl, A., Roth, C., Gressel, S., Baejen, C., Soeding, J., and Cramer, P. (2019). Transcriptome maps of general eukaryotic RNA degradation factors. *eLife*, 8:e47040.

Sohrabi-Jahromi, S. and Söding, J. (2021). Thermodynamic modeling reveals widespread multivalent binding by RNA-binding proteins. *bioRxiv*.

Spitzer, J., Hafner, M., Landthaler, M., Ascano, M., Farazi, T., Wardle, G., Nusbaum, J., Khorshid, M., Burger, L., Zavolan, M., and Tuschl, T. (2014). Chapter Eight - PAR-CLIP (Photoactivatable Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation): a Step-By-Step Protocol to the Transcriptome-Wide Identification of Binding Sites of RNA-Binding Proteins. In Lorsch, J., editor, *Laboratory Methods in Enzymology: Protein Part B*, volume 539 of *Methods in Enzymology*, pages 113–161. Academic Press.

Sprinzl, M., Horn, C., Brown, M., Ioudovitch, A., and Steinberg, S. (1998). Compilation of tRNA sequences and sequences of tRNA genes. *Nucleic acids research*, 26(1):148–153.

Standart, N. and Weil, D. (2018). P-Bodies: cytosolic droplets for coordinated mRNA storage. *Trends in Genetics*, 34(8):612–626.

Steiger, M., Carr-Schmid, A., Schwartz, D. C., Kiledjian, M., and Parker, R. (2003). Analysis of recombinant yeast decapping enzyme. *RNA*, 9(2):231–238.

Sternburg, E. L. and Karginov, F. V. (2020). Global Approaches in Studying RNA-Binding Protein Interaction Networks. *Trends in Biochemical Sciences*, 45(7):593–603.

Stevens, A. (2001). 5'-Exoribonuclease 1: Xrn1. *Methods in Enzymology*, 342:251–259.

Stewart, M. (2010). Nuclear export of mRNA. *Trends in Biochemical Sciences*, 35(11):609 – 617.

Stewart, M. (2019). Polyadenylation and nuclear export of mRNAs. *Journal of Biological Chemistry*, 294(9):2977 – 2987.

Stitzinger, S. H., Sohrabi-Jahromi, S., and Söding, J. (2021). Cooperativity boosts affinity and specificity of proteins with multiple RNA-binding domains. *bioRxiv*.

Stowell, J. A., Webster, M. W., Kögel, A., Wolf, J., Shelley, K. L., and Passmore, L. A. (2016). Reconstitution of targeted deadenylation by the Ccr4-Not complex and the YTH domain protein Mmi1. *Cell Reports*, 17(8):1978–1989.

Stražar, M., Žitnik, M., Zupan, B., Ule, J., and Curk, T. (2016). Orthogonal matrix factorization enables integrative analysis of multiple RNA binding proteins. *Bioinformatics*, 32(10):1527–1535.

Strom, A. R. and Brangwynne, C. P. (2019). The liquid nucleome – phase transitions in the nucleus at a glance. *Journal of Cell Science*, 132(22).

Su, Y., Luo, Y., Zhao, X., Liu, Y., and Peng, J. (2019). Integrating thermodynamic and sequence contexts improves protein-RNA binding prediction. *PLOS Computational Biology*, 15(9):1–14.

Sun, T., Li, Q., Xu, Y., Zhang, Z., Lai, L., and Pei, J. (2019). Prediction of liquid-liquid phase separation proteins using machine learning. *bioRxiv*.

Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic Attribution for Deep Networks. In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328, International Convention Centre, Sydney, Australia. PMLR.

Sweet, T., Kovalak, C., and Coller, J. (2012). The DEAD-box protein Dhh1 promotes decapping by slowing ribosome movement. *PLoS Biol*, 10(6):e1001342.

Talkish, J., May, G., Lin, Y., Woolford, J. L., and McManus, C. J. (2014). Mod-seq: high-throughput sequencing for chemical probing of RNA structure. *RNA*, 20(5):713–720.

Teixeira, D. and Parker, R. (2007). Analysis of P-body assembly in Saccharomyces cerevisiae. *Molecular biology of the cell*, 18(6):2274–2287.

Thiebaut, M., Kisseleva-Romanova, E., Rougemaille, M., Boulay, J., and Libri, D. (2006). Transcription Termination and Nuclear Degradation of Cryptic Unstable Transcripts: A Role for the Nrd1-Nab3 Pathway in Genome Surveillance. *Molecular Cell*, 23(6):853–864.

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

Tsai, M.-C., Manor, O., Wan, Y., Mosammaparast, N., Wang, J. K., Lan, F., Shi, Y., Segal, E., and Chang, H. Y. (2010). Long Noncoding RNA as Modular Scaffold of Histone Modification Complexes. *Science*, 329(5992):689–693.

Tuck, A. C., Rankova, A., Arpat, A. B., Liechti, L. A., Hess, D., Iesmantavicius, V., Castelo-Szekely, V., Gatfield, D., and BÃŒhler, M. (2020). Mammalian RNA Decay Pathways Are Highly Specialized and Widely Linked to Translation. *Molecular Cell*, 77(6):1222–1236.e13.

Tucker, M., Staples, R. R., Valencia-Sanchez, M. A., Muhlrad, D., and Parker, R. (2002). Ccr4p is the catalytic subunit of a Ccr4p/Pop2p/Notp mRNA deadenylase complex in Saccharomyces cerevisiae. *The EMBO Journal*, 21(6):1427–1436.

Tudek, A., Porrua, O., Kabzinski, T., Lidschreiber, M., Kubicek, K., Fortova, A., Lacroute, F., Vanacova, S., Cramer, P., Stefl, R., and Libri, D. (2014). Molecular Basis for Coordinating Transcription Termination with Noncoding RNA Degradation. *Molecular Cell*, 55(3):467–481.

Uhl, M., Houwaart, T., Corrado, G., Wright, P. R., and Backofen, R. (2017). Computational analysis of CLIP-seq data. *Methods*, 118-119:60–72. Protein-RNA: Structure Function and Recognition.

Uhler, J. P., Hertel, C., and Svejstrup, J. Q. (2007). A role for noncoding transcription in activation of the yeast PHO5 gene. *Proceedings of the National Academy of Sciences*, 104(19):8011–8016.

van Dijk, E., Cougot, N., Meyer, S., Babajko, S., Wahle, E., and Seraphin, B. (2002). Human Dcp2: a catalytically active mRNA decapping enzyme located in specific cytoplasmic structures. *The EMBO Journal*, 21(24):6915–6924.

Van Driesche, S. J. and Martin, K. C. (2018). New frontiers in RNA transport and local translation in neurons. *Developmental Neurobiology*, 78(3):331–339.

van Hoof, A., Frischmeyer, P. A., Dietz, H. C., and Parker, R. (2002). Exosome-Mediated Recognition and Degradation of mRNAs Lacking a Termination Codon. *Science*, 295(5563):2262–2264.

van Mierlo, G., Jansen, J. R., Wang, J., Poser, I., van Heeringen, S. J., and Vermeulen, M. (2021). Predicting protein condensate formation using machine learning. *Cell Reports*, 34(5):108705.

Van Nostrand, E. L., Pratt, G. A., Shishkin, A. A., Gelboin-Burkhart, C., Fang, M. Y., Sundararaman, B., Blue, S. M., Nguyen, T. B., Surka, C., Elkins, K., et al. (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature Methods*, 13(6):508–514.

Vaňáčová, Š., Wolf, J., Martin, G., Blank, D., Dettwiler, S., Friedlein, A., Langen, H., Keith, G., and Keller, W. (2005). A new yeast poly (A) polymerase complex involved in RNA quality control. *PLoS Biol*, 3(6):e189.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., et al. (2001). The sequence of the human genome. *science*, 291(5507):1304–1351.

Vernon, R. M., Chong, P. A., Tsang, B., Kim, T. H., Bah, A., Farber, P., Lin, H., and Forman-Kay, J. D. (2018). Pi-Pi contacts are an overlooked protein feature relevant to phase separation. *eLife*, 7:e31486.

Vernon, R. M. and Forman-Kay, J. D. (2019). First-generation predictors of biological protein phase separation. *Current opinion in structural biology*, 58:88–96.

Vicens, Q., Kieft, J. S., and Rissland, O. S. (2018). Revisiting the closed-loop model and the nature of mRNA 5'–3' communication. *Molecular Cell*, 72(5):805–812.

Wan, Y., Kertesz, M., Spitale, R. C., Segal, E., and Chang, H. Y. (2011). Understanding the transcriptome through RNA structure. *Nature Reviews Genetics*, 12(9):641–655.

Wang, J., Choi, J.-M., Holehouse, A. S., Lee, H. O., Zhang, X., Jahnel, M., Maharana, S., Lemaitre, R., Pozniakovsky, A., Drechsel, D., et al. (2018). A molecular grammar governing the driving forces for phase separation of prion-like RNA binding proteins. *Cell*, 174(3):688–699.

Wang, L., Lewis, M. S., and Johnson, A. W. (2005). Domain interactions within the Ski2/3/8 complex and between the Ski complex and Ski7p. *RNA*, 11(8):1291–1302.

Warner, J. R. (1999). The economics of ribosome biosynthesis in yeast. *Trends in biochemical sciences*, 24(11):437–440.

Wells, S. E., Hillner, P. E., Vale, R. D., and Sachs, A. B. (1998). Circularization of mrna by eukaryotic translation initiation factors. *Molecular Cell*, 2(1):135 – 140.

Wikipedia (2021). RNA world — Wikipedia, The Free Encyclopedia. `https://en.wikipedia.org/w/index.php?title=RNA_world&oldid=1006130483`. [Online; accessed 01-February-2021].

Will, C. L. and Lührmann, R. (2011). Spliceosome structure and function. *Cold Spring Harbor Perspectives in Biology*, 3(7):a003707.

Wyers, F., Rougemaille, M., Badis, G., Rousselle, J.-C., Dufour, M.-E., Boulay, J., Regnault, B., Devaux, F., Namane, A., Seraphin, B., Libri, D., and Jacquier, A. (2005). Cryptic Pol II Transcripts Are Degraded by a Nuclear Quality Control Pathway Involving a New Poly(A) Polymerase. *Cell*, 121(5):725–737.

Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Münster, S., Camblong, J., Guffanti, E., Stutz, F., Huber, W., and Steinmetz, L. M. (2009). Bidirectional promoters generate pervasive transcription in yeast. *Nature*, 457(7232):1033–1037.

Yan, J. and Zhu, M. (2020). A Review About RNA-Protein-Binding Sites Prediction Based on Deep Learning. *IEEE Access*, 8:150929–150944.

Youn, J.-Y., Dyakov, B. J., Zhang, J., Knight, J. D., Vernon, R. M., Forman-Kay, J. D., and Gingras, A.-C. (2019). Properties of stress granule and P-body proteomes. *Molecular cell*, 76(2):286–294.

Yu, H., Wang, J., Sheng, Q., Liu, Q., and Shyr, Y. (2018). beRBP: binding estimation for human RNA-binding proteins. *Nucleic Acids Research*, 47(5):e26–e26.

Zappulla, D. C. and Cech, T. R. (2004). Yeast telomerase RNA: A flexible scaffold for protein subunits. *Proceedings of the National Academy of Sciences*, 101(27):10024–10029.

Zeng, M., Chen, X., Guan, D., Xu, J., Wu, H., Tong, P., and Zhang, M. (2018). Reconstituted post-synaptic density as a molecular platform for understanding synapse formation and plasticity. *Cell*, 174(5):1172–1187.

Zhang, B. and Herman, P. (2020). It is all about the process (ing): P-body granules and the regulation of signal transduction. *Current Genetics*, 66(1):73–77.

Zhang, S., Zhou, J., Hu, H., Gong, H., Chen, L., Cheng, C., and Zeng, J. (2016). A deep learning framework for modeling structural features of RNA-binding protein targets. *Nucleic Acids Research*, 44(4):e32–e32.

Appendices

A1 BMF User Guide

A1.1 Contents

A1.2 Summary

Knowing the basis of protein-RNA recognition is essential to understanding regulatory processes in the cell. Many RNA-binding proteins (RBPs) form complexes or have multiple domains that allow binding to the RNA molecule in a multivalent manner. Through cooperative binding these proteins can reach higher specificity and affinity than those of single RNA-binding domains. However, current approaches to RNA de novo motif discovery do not take the modularity of binding events into account. Here we present Bipartite Motif Finder (BMF), an RNA motif finder that is based on a thermodynamic model of an RBP with two binding sites acting cooperatively in targeting an RNA molecule. We show that bipartite binding is a common strategy among RBPs to achieve higher levels of sequence specificity. We furthermore illustrate that the spacial geometry between the two binding sites can be learnt from bound RNA sequences and that this information enhances the model's accuracy in predicting new binding sites. These bipartite motifs are consistent with previously known motifs and binding behaviors. Our results demonstrate the importance of multivalent binding for RNA-binding proteins and highlight the value of bipartite motif models in representing the multivalency of protein-RNA interactions.

BMF is also available as a webserver:

- Link: bmf.soedinglab.org
- Web server repository: soedinglab/bmf-webserver

A1.3 Installation

Requirements

- `python>3.6`

- `numpy`
- `cython`

Installing requirements with Conda: Create a new conda environment with `python`, `numpy`, and `cython`:

```
conda create -n bmf python=3.6 numpy cython
conda activate bmf
```

```
sudo apt-get update
sudo apt-get install python3.6 python3-pip
pip3 install numpy cython
```

Installing requirements on Ubuntu without Conda:

```
brew install python3
pip install numpy cython
```

Installing requirements on MacOS with brew:

BMF installation

1. **Optional:** BMF is also available as a faster version for running on AVX2 extension capable processor. You can check if AVX2 is supported by executing `cat /proc/cpuinfo | grep avx2` on Linux and `sysctl -a | grep machdep.cpu.leaf7_features | grep AVX2` on MacOS). If your processor supports AVX2, run the following command to compile a faster version of BMF:

```
export USE_AVX=1
```

2. Install BMF with pip:

```
pip install https://github.com/soedinglab/bipartite_motif_finder/releases/download/
v1.0.0a/bmf_tool-1.0.0.tar.gz
```

See BMF help page:

```
bmf --help
```

A1.4 BMF guide

BMF has three main functionalities: (1) learning *de novo* bipartite motifs from enriched and background sequence sets, (2) plotting the motif and predicting if the RNA-binding protein has a bipartite motif or not, and (3) using the trained BMF model to predict binding to new sequences.

In the following sections, we describe how to use BMF to perform each of these functionalities.

A1.5 Motif discovery

You can call the command-line tool `bmf` to perform *de novo* motif discovery. Here is a list of parameters that you can pass to `bmf` for training:

```
positional arguments:

  sequences             path to positive sequences enriched with the
                        motif.

compulsory arguments:

  --BGsequences BGSEQUENCES
                        path to background sequences.

optional arguments:

  --input_type {fasta,fastq,seq}
                        format of input sequences. Can be "fasta", "fastq",
                        or "seq". [Default:"fasta"]

  --motif_length MOTIF_LENGTH
                        the length of each core in the bipartite motif.
                        [Default:3]

  --no_tries NO_TRIES   the number of times the program is run with random
                        initializations.
                        [Default:5]

  --output_prefix OUTPUT_PREFIX
                        output file prefix. You can specify a directory e.g.
                        "--output_prefix output_dir/my_prefix"
                        [Default:"bipartite"]
```

```
--var_thr VAR_THR     variability threshold condition to stop ADAM
                      [Default:0.03]

--batch_size BATCH_SIZE
                      the number of sequences processed in each batch of stochastic
                      gradient descent.
                      [Default:512]

--max_iterations MAX_ITERATIONS
                      max number of iterations before stopping ADAM.
                      [Default:1000]

--no_cores NO_CORES   the numbers of CPU cores used
                      [Default:4]
```

Run BMF with multiple random initializations

You can run BMF with with `n` random parameter initializations by specifying `--no_tries`. Even though BMF is robust to parameter initializations, this ensures that the best likelihood model would be found. In our manuscript we run BMF with `--no_tries`. We develop the BMF workflow in a way that when multiple initializations are performed, the best likelihood solution will be used to generate the sequence logo and to predict binding.

Output file name

You can specify the output file name with `--output_prefix path-to-file/file-name`. BMF will generate the following outputs for each round of parameter initialization `i` (`i` is between 1 and `n` for `n` initializations):

- Plots of parameter changes over iterations and training set ROC curve: `path-to-file/file-name_cs{motif_length}_{i}.pdf & .png`
- Model parameters `path-to-file/file-name_cs{motif_length}_{i}.txt`

Input file formats

You can run BMF with traditional "fasta" and "fastq" file formats. Additionally you can provide just the sequences in the following format which we refer to as "seq":

```
AGGCTCGGTTACGTGCAGGGCCTGATGTTCTTGATCTGTT
CTTCCAAGGAAGCTTTGACTCACAGAAATGGTAAAGTCCA
TCCCTTCGCTAAGTAGGGACGCCTCGGGCGAGACAATAGC
GAGGTGGGCTCGCGTACCTCACTTACACCATGCGCCTCAT
...
```

Note: The input sequences **should** be of equal lengths, and can only consist of the characters: A, C, G, T, U, and N.

A1.6 Generate motif logo

You can generate bmf logo plots, using the parameter files generated via `bmf` in the previous step. To do so you need to call `bmf_logo` with the following parameters:

```
positional arguments:
  parameter_prefix      path-to-bmf-param-file that specifies model parameters or
                        when multiple parameters exist, their common root.

optional arguments:
  --motif_length MOTIF_LENGTH
                        the length of each core in the bipartite motif
                        [Default:3]
```

Please note that `parameter_prefix` corresponds to `output_prefix` in the previous step. When multiple initializations were used, `bmf_logo` reads all and selects the best likelihood solution to generate the motif logo.

The BMF logo plot is stored at `{parameter_prefix}_seqLogo.pdf & .png`.

A1.7 Predict binding

You can use the trained BMF model parameters to predict binding scores for new sequences. To do so you should run `bmf` with `--predict`. Here is a list of parameters that you can pass to `bmf` for predicting:

```
positional arguments:

  sequences             path to test sequences.

compulsory arguments:

  --test

  --model_parameters MODEL_PARAMETERS
                        path to .txt file that specifies model parameters,
                        or the output_prefix used when training bmf.

optional arguments:

  --input_type {fasta,fastq,seq}
```

```
                      format of input sequences. Can be "fasta", "fastq",
                      or "seq".
                      [Default:"fasta"]

  --motif_length MOTIF_LENGTH
                      the length of each core in the bipartite motif.
                      [Default:3]
```

The binding score for each sequence is saved in the file `{model_parameters}.predictions`. **Note:** these values correspond to the summation of statistical weights over all possible configurations. Higher values correspond to a higher binding probability. Based on our thermodynamic model, these values can be converted to binding probabilities with the following formula:

A1.8 Example workflow

You can find the fasta files needed to run this example in `data` directory. Here we run BMF with one random parameter initialization. You can change the `--no_tries` to increase the number of BMF runs with new initial parameter values. The best likelihood solution would be used in this case to plot the BMF logo, and to predict binding to new sequences.

Motif discovery

You can use `bmf` in training mode for *de novo* motif discovery. By default, BMF runs over a maximum of 1000 iterations.

```
bmf positives_AAA_CCC.fasta --BGsequences negatives_AAA_CCC.fasta --input_type fasta
--output_prefix AAA_CCC --motif_length 3  --no_tries 1
```

Generate sequence logo

You can use `bmf_logo` to plot the best likelihood motif model generated by BMF. Specify the `output_prefix` from the previous step to allow `bmf_logo` to find all associated parameter files. Here we use `AAA_CCC` to specify the outputs from the previous run:

```
bmf_logo AAA_CCC --motif_length 3
```

Predict binding to new sequences

You can use the trained BMF model parameters to predict binding scores for new sequences. To specify `--model_parameters`, use the `output_prefix` from the first step (here `AAA_CCC`).

```
bmf test_sequences.fasta  --predict --input_type fasta --model_parameters AAA_CCC
--output_prefix predict_test_sequences
```

A1.9 License terms

The software is made available under the terms of the GNU General Public License v3.