COMPUTATIONAL INVESTIGATION OF PROTEIN-RNA INTERACTIONS DETECTED BY CLIP, THEIR SPECIFICITY AND DYNAMICS IN EMBRYONIC DEVELOPMENT

Klara Kuret

Doctoral Dissertation
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Supervisor: Dr. Jernej Ule,
[1]National Institute of Chemistry, Hajdrihova 19, SI-1001, Ljubljana, Slovenia
[2]Dementia Research Institute at KCL, London, UK
Co-Supervisor: Dr. Miha Modic,
[1]National Institute of Chemistry, Hajdrihova 19, SI-1001, Ljubljana, Slovenia
[2]Dementia Research Institute at KCL, London, UK

Evaluation Board:
Prof. Tamara Lah Turnšek, Chair,
[1]National Institute of Biology, Ljubljana, Slovenia
Prof. Annalisa Marsico, Member,
[1]Computational Health Center, Helmholtz Center Munich, Munich, Germany
Prof. Nejc Haberman, Member,
[1]Institute of Clinical Sciences, Imperial College London, Hammersmith Hospital Campus, London, UK.

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Klara Kuret

COMPUTATIONAL INVESTIGATION OF PROTEIN-RNA INTERACTIONS DETECTED BY CLIP, THEIR SPECIFICITY AND DYNAMICS IN EMBRYONIC DEVELOPMENT

Doctoral Dissertation

UPORABA RAČUNSKIH PRISTOPOV ZA PREUČEVANJE STIKOV MED PROTEINI IN RNK, ZAZNANIH Z METODO CLIP, TER NJIHOVE SPECIFIČNOSTI IN DINAMIKE V EMBRIONALNEM RAZVOJU

Doktorska disertacija

Supervisor: Dr. Jernej Ule

Co-Supervisor: Dr. Miha Modic

Ljubljana, Slovenia, December 2023

To delo posvečam mami Katarini in očetu Zoranu.

Acknowledgments

I am deeply grateful to everyone who stood by me during the creation of this work—there have been so many of you and I am blessed to have each one of you in my life.

First and foremost, I want to thank my mentor Jernej Ule who is an endless source of inspiration and kindness. You are a one-of-a-kind scientist and I aspire to be more like you. Your calm guidance and knowledge were invaluable to this work and have shaped me as a person. Thank you for starting a lab in Ljubljana, which offers young scientists the opportunity to conduct breakthrough research and connects them to leading scientists worldwide.

To Miha Modic—the exceptional and endlessly curious scientist. You gave me so much knowledge, but, most importantly, you taught me patience and persistence.

To Aram Amalietti—if it was not for you I might not have found my way here. You taught me Python and data science, before we even had an office. It was not always easy watching me struggle with code, but you did it anyway—albeit with a grunt here or there. Thank you.

To Miha Milek—my mentor during my first year. You built this lab from the ground up. I am not afraid to say this on behalf of everyone who works here now—thank you! We miss your positive and relaxed company.

To my colleagues Tajda, Anja, Urška, Maks, Aniela, and Jona—you inspired me with your outstanding work, pushed me to be better, and created a warm and cooperative lab environment that is difficult to come by. I am so proud of how we have grown together.

To my partner Nejc—you are my anchor, my strength, my energy. Your love made this easy.

To my parents Katarina and Zoran, thank you for your unwavering support and your faith in me.

To my brother Klemen, sister Živa and to my friends Teja and Maša—thank you for the moments of laughter and light-hearted fun during these years.

To the evaluation committee that reviewed this thesis—your time, effort, and professional advice enabled me to improve this work. Thank you.

Finally, I thank the National Institute of Chemistry and the Francis Crick Institute that supported this work. This research was funded by the European Union's Horizon 2020 research and innovation programme (835300-RNPdynamics).

Abstract

RNA molecules dynamically interact with RNA-binding proteins (RBPs), which control various aspects of RNA fate, such as its processing, localisation, and stability. Intricate networks of protein-RNA interactions thereby regulate gene expression and have a profound effect on downstream cellular processes. Most RBPs recognise specific motifs on their bound RNAs, characterised by linear nucleotide sequences, RNA structures, RNA modifications, or combinatorial patterns of these features. By understanding how these features determine the specificity of protein-RNA interactions, we can gain valuable insights into the fundamental mechanisms that govern gene expression and cellular processes. As the first step, protein-RNA interactions are identified within cells by transcriptomic experiments such as crosslinking and immunoprecipitation (CLIP), which identifies RNA crosslink sites of selected RBPs. However, each method has its unique experimental biases, which creates challenges in separating biological from technical signal. For this purpose, we introduce the positionally-enriched k-mer analysis (PEKA), a computational approach for discovery of linear-sequence motifs enriched in the proximity of RNA crosslink sites identified by CLIP. PEKA implements steps to minimise the effects of technical biases in motif discovery and by visualising the motifs around crosslink sites facilitates the comparison between distinct CLIP datasets. Here, we apply PEKA for a comparative study of binding motifs that are enriched across an array of distinct RBPs and CLIP methods, to gain insights into general features underlying RBP specificity, such as the presence of intrinsically disordered regions (IDRs) and canonical RNA-binding domains, as well as the variations in technical artefacts and specificity between CLIP datasets.

Furthermore, we extended the application of PEKA and related computational approaches to investigate the dynamic regulation of LIN28A—an RBP essential for promoting the switch from naïve to primed cell fate in early embryonic development. In this process, LIN28A mediates the rapid decay of naïve-pluripotency factor mRNAs, but it was unclear how it selectively targets these mRNAs. Our findings show that selectivity is achieved by activating LIN28A through phosphorylation of its IDR, profoundly changing its RNA interactions. Upon phosphorylation, LIN28A converges to AU-rich 3′-UTR termini bound by the cytoplasmic poly(A)-binding proteins. mRNAs targeted for decay exhibited higher multivalency of AU-rich motifs and a greater accumulation of both LIN28A and poly(A)-binding proteins at their terminal regions. Given that dysregulation of LIN28A is linked to cancers and diverse growth conditions, a deeper understanding of its regulatory mechanisms is crucial. This work adds to that understanding and demonstrates the value of comparative CLIP analyses to elucidate methodological biases, determinants of RBP-binding specificity and gain functional insights into how RBPs regulate the specificity and dynamics of cellular processes.

Povzetek

RNK-vezavni proteini (RBP-ji) so ključni regulatorji izražanja genov v celici. Sodelujejo pri raznovrstnih procesih metabolizma RNK molekul in tako omogočajo ustrezno delovanje celic in njihov odziv na signale iz okolja. Posamezen protein, ki se veže na RNK molekule, na njih prepozna specifične vezavne motive, ki jih definirajo nukleotidno zaporedje RNK, struktura RNK molekul, post-transkripcijske modifikacije nukleotidov ali kombinatorični vzorci teh značilnosti. Z razumevanjem, kako te značilnosti določajo specifičnost interakcij med proteini in RNK, lahko pridobimo vpogled v temeljne mehanizme, ki nadzirajo izražanje genov in celične procese. Stike med proteini in RNK v celicah zaznamo z metodo premreženja in imunoprecipitacije (CLIP), ki identificira mesta premreženja specifičnih RBP-jev z RNK na ravni celotnega transkriptoma. Kljub razširjeni uporabi metode CLIP, učinkovita karakterizacija interakcij med proteini in RNK ostaja izziv, saj je zaradi nespecifičnega ozadja metode otežena zaznava bioloških signalov, ki predstavljajo vezavo tarčnega RBP-ja. V sklopu tega dela smo razvili računalniški pristop PEKA, ki identificira nukleotidne motive, obogatene v bližini mest premreženja na RNK, zaznanih z metodo CLIP. PEKA implementira pristope za zmanjšanje vpliva metodološkega ozadja na odkrivanje vezavnih motivov ter omogoča poglobljeno analizo motivov znotraj posameznih transkriptomskih regij in njihovega pozicioniranja okoli mest premreženja. PEKO smo uporabili za primerjavo vezavnih motivov, ki so obogateni v različnih RBP-jih in metodah CLIP. Ta raziskava nam je omogočila razumevanje lastnosti, ki vplivajo na specifičnost RBP-jev, kot so vsebnost intrinzično neurejenih regij in kanoničnih RNK-vezavnih domen; kot tudi razumevanje vpliva metodološkega ozadja na zaznane motive.

Poleg tega smo PEKO aplicirali skupaj s komplementarnimi računalniškimi pristopi za analizo dinamične regulacije RBP-ja LIN28A, ki koordinira celično diferenciacijo iz naivnega v napredno prehodno pluripotentno celično stanje (ang. *primed state*) v zgodnjem embrionalnem razvoju. V tem procesu LIN28A regulira hiter in selektiven razpad mRNK molekul, ki kodirajo proteine, nujne za vzdrževanje naivne pluripotence. Naša raziskava je pokazala, da se selektivnost razkroja mRNK doseže s fosforilacijo LIN28A v intrinzično neurejeni regiji, kar spremeni njegove interakcije z RNK tarčami. Po fosforilaciji se LIN28A nakopiči na 3′-koncih mRNK, ki so bogati z AU-motivi in na katere so predhodno vezani citoplazemski poli(A)-vezavni proteini. mRNK, ki se v prehodu iz naivnega v *primed* celično stanje razgradijo, imajo višjo multivalentnost AU-motivov in povečano vezavo LIN28A in poli(A)-vezavnih proteinov v svojih 3′-končnih regijah. Ker je disregulacija LIN28A povezana z rakavimi obolenji in z drugimi kompleksnimi boleznimi, je globlje poznavanje njegovih regulatornih mehanizmov ključnega pomena za razumevanje teh bolezni in osnovanje ustreznih terapevtskih pristopov. To doktorsko delo karakterizira še nepoznan regulatorni mehanizem LIN28A in pokaže uporabno vrednost primerjalnih analiz podatkov CLIP za širše razumevanje metodoloških ter bioloških vplivov na zaznano specifičnost vezave RBP-jev, kot tudi za analizo funkcije RBP-jev v celičnih procesih.

Contents

List of Figures

List of Tables

Abbreviations

CDS	... Coding sequence
CLIP	... Crosslinking and immunoprecipitation
CSD	... Cold-shock domain
cDNA	... Complementary DNA (DNA obtained in the process of reverse transcription that is complementary to template RNA)
ENCODE	... Encyclopaedia of DNA elements
eRIC	... Enhanced RNA-interactome capture
IDR	... Intrinsically disordered region
KH	... K-Homology domain
miRNA	... Micro RNA
nt	... nucleotide
oXn	... Out-of-peak crosslink (represents background signal in PEKA)
PABP	... Poly(A)-binding protein
PAS	... Poly(A) signal
PCR	... Polymerase chain reaction
PEKA	... Positionally-enriched k-mer analysis
RBD	... RNA-binding domain
RBNS	... RNA-Bind-N-Seq
RBP	... RNA-binding protein
RNA	... Ribonucleic acid
RRM	... RNA-recognition motif
RT	... Reverse transcription
SMInput	... Size-matched input
tXn	... Thresholded crosslink (represents foreground signal in PEKA)
UTR	... Untranslated region
Xn	... crosslink
ZnF	... Zinc-finger domain

Glossary

5′-end	...	The 5′-end of an RNA molecule refers to the end of the RNA strand that has the fifth carbon in the sugar-ring of the ribose at its terminus.
3′-end	...	The 3′-end of an RNA molecule refers to the end of the RNA strand that has the third carbon in the sugar-ring of the ribose at its terminus.
Poly(A) signal	...	Poly(A) signal is a specific sequence of nucleotides in the DNA—usually AATAAA—that signals the addition of a poly(A) tail during the process of mRNA synthesis.
Crosslink	...	Biologically, the term crosslink in CLIP experiment refers to a single-nucleotide position on the RNA, which forms a covalent bond with the protein upon UV-crosslinking. In bioinformatic analysis of CLIP data, however, the term crosslink refers to a diagnostic event that identifies a protein-RNA interaction, like a cDNA truncation or mutation that occurs during reverse transcription of crosslinked RNA fragment into cDNA.

Chapter 1

Introduction

The aim of this thesis is to increase our understanding of specificity underlying protein-RNA interactions and their role in regulation of cellular processes. The main contributions include the development of a computational tool for effective discovery of sequence motifs recognised by RNA-binding proteins (RBPs) from crosslinking and immunoprecipitation (CLIP) data, the identification of general RBP features that influence the specificity of their interactions with RNA, an increased understanding of experimental biases in CLIP and their impact on detected binding motifs. Furthermore, this work reveals novel insights into the mechanism by which LIN28A regulates selective mRNA decay in embryogenesis. In this introductory chapter, we present the research background, motivate the problems addressed and overview solutions to the presented problems. Next, we state the purposes of the dissertation, its goals, and scientific contributions. We conclude with a structural overview of the rest of the thesis.

1.1 Background

In the following sections, we present the state-of-the-art research related to the formation of protein-RNA interactions, experimental methods of their detection *in vivo* and *in vitro,* computational workflows for the analysis of CLIP data, and approaches for characterising the specificity of detected protein-RNA interactions. We also describe the documented roles of LIN28A, an essential RBP with diverse functions in embryonic development and other growth-related processes.

1.1.1 Specificity of protein-RNA interactions

RBPs represent one of the most abundant protein-classes in the cell, with over 1500 proteins representing 7.5% of all protein-coding genes in the human genome (Gerstberger et al., 2014). RBPs are evolutionarily deeply conserved and generally ubiquitously expressed, which speaks to their central and conserved role in gene regulation (Gerstberger et al., 2014). RBP-RNA interactions mediate a wide array of processes related to RNA biogenesis, localisation, translation, and more (Gebauer et al., 2020). By identifying the RBP-RNA interactions in cells and characterising their defining features, we can obtain valuable insights into the regulatory mechanism of a specific RBP. The features that influence the specificity of protein-RNA interactions can be intrinsic to the RNA molecules or extrinsic, i.e., influenced by other factors in the cellular environment, rather than intrinsic to the molecules of RNA and protein that interact. Intrinsic RNA features include the nucleotide sequence, phosphate backbone, as well as local and distal RNA structures. In contrast, extrinsic features include post-translational RBP modifications, post-

transcriptional modifications on the RNA, as well as respective expression and localisation of RBPs and RNAs (Dasti et al., 2020).

Following the structure-function paradigm, RNA-binding domains (RBDs) and their combinations enable RBPs to recognise distinct binding sites on the RNA. Canonical RBDs, such as RNA-recognition motif (RRM), K-Homology domain (KH), CCCH zinc finger, and Pumillio domains, are associated with the recognition of specific nucleotide sequences (Auweter et al., 2006; Fu & Blackshear, 2017; Maris et al., 2005; Nicastro et al., 2015). These domains diversified early in the process of evolution and are very evolutionarily conserved, however RBPs containing these types of domains only represent a minority of all proteins that associate with the RNA in the cell. The other class of RBPs primarily recognises RNA structure and/or RNA backbone interactions, with zinc-finger domains and double-stranded RNA-binding domains. Proteins in this class include, for example, RNA helicases and RNA exonucleases. Beyond sequence and structure specific RBPs, many other proteins also associate with the RNA, which do not contain any of the known RBDs. For these unconventional RBPs little is known about their specificity toward RNA and about the way they form RNA interactions.

To explore the features that RBPs recognise on the RNA *in vivo*, the common starting point is to determine precise locations of RBP-RNA contacts on the transcriptome with CLIP methods (F. C. Y. Lee & Ule, 2018). Commonly used approaches for high-throughput profiling of RBP specificity also include *in vitro* approaches, in which recombinant RBP is used to enrich RNA ligands from a random pool of RNA sequences.

In the following sections, we describe the methodology of CLIP and *in vitro* selectivity assays, as well as computational analysis of respective data, and the process of motif-discovery, which is used to characterise the binding motifs of RBPs.

1.1.1.1 CLIP methods

CLIP methods are widely used to determine RBP binding sites *in vivo*, as they can identify the locations of RNA-protein contacts with near-nucleotide resolution (Hafner et al., 2021). They achieve this by inducing the formation of covalent bonds between protein and RNA molecules in direct proximity (i.e. crosslinking), thus stabilising the RNPs present in the sample (Hafner et al., 2021; F. C. Y. Lee & Ule, 2018; Ule et al., 2003). After crosslinking, RNPs are released from the cells by lysis and treated with RNase to partially digest the RNA which is not protected by RBP binding. Afterwards, the target RBP is immunopurified along with the crosslinked RNA fragments and the isolate is separated by SDS-PAGE to ensure that only the RNA fragments crosslinked to the protein of interest are isolated. The RNA fragments are then excised from the gel, reversed transcribed into complementary DNA (cDNA), and amplified by polymerase-chain reaction (PCR) to prepare cDNA libraries for high-throughput sequencing (Hafner et al., 2021; F. C. Y. Lee & Ule, 2018). The resulting sequencing data is then processed with computational methods to determine precise locations of protein-RNA contacts and characterise RBP binding sites (Hafner et al., 2021). Multiple CLIP protocols have developed over time; however, all techniques still maintain the same core stages of crosslinking, immunoprecipitation, and library preparation (Lee and Ule, 2018). In this work, we utilise data produced by three broadly adopted CLIP methods, which will be presented in the following paragraphs: individual-nucleotide resolution CLIP (iCLIP) (König et al., 2010), enhanced CLIP (eCLIP) (Van Nostrand et al., 2016) and photoactivatable ribonucleoside-enhanced CLIP (PAR-CLIP) (Garzia et al., 2017; Hafner et al., 2010). For comprehensive information on other CLIP protocols, we refer the reader to the following review (F. C. Y. Lee & Ule, 2018).

1.1.1.1.1 iCLIP

The original CLIP protocol employed Sanger sequencing to determine the complete RNA fragments crosslinked to a specific RBP (Ule et al., 2003). With advancements in technology, CLIP methods were subsequently enhanced by integrating them with high-throughput sequencing, which markedly improved their sensitivity and resolution (Chi et al., 2009; Licatalosi et al., 2008; Yeo et al., 2009). Among these innovations, iCLIP emerged as a pioneering CLIP variant capable of identifying the precise nucleotide positions of protein-RNA contacts at a large scale (König et al., 2010). Owing to its exceptional sensitivity and high resolution, iCLIP and its variants continue to be among the most extensively utilised CLIP protocols in contemporary scientific research.

iCLIP protocol uses irradiation with UVC light (wavelength of 254 nm) to induce crosslinking of nucleotides, with proximal reactive amino acid side chains (König et al., 2010). To identify the precise locations of crosslink sites, the method exploits the property of reverse transcriptase to frequently terminate at the crosslinked nucleotide, due to the remnants of crosslinked peptide which presents a steric obstacle in the path of the enzyme. Premature termination of reverse transcription (RT) at the crosslink site results in a library of truncated cDNA fragments with a 5′-end mapping to the first nucleotide downstream of the crosslink (König et al., 2010), which enables identification of crosslink sites with near-nucleotide-level resolution. In some cases, however, reverse transcriptase transcribes the entire read, resulting in read-through cDNAs that are not truncated and thus not indicative of the actual crosslink sites. To account for these instances, some studies appropriately refer to positions identified by 5′-ends of reads as *diagnostic crosslinking events* (Bahrami-Samani et al., 2015; Feng et al., 2019; A. Shah et al., 2017), however, we will simply refer to these sites as *crosslinks*.

After crosslinking, cells are lysed using a lysis buffer and subjected to sonication to extract protein-RNA complexes in a clear lysate. Next, the RNA is partially digested with RNase I to generate fragments within the optimal size range of 30 to 200 nucleotides (nts) (F. C. Y. Lee & Ule, 2018). Unlike other RNases, RNase I exhibits uniform cleavage activity across all nucleotides, minimising the sequence bias of RNA fragmentation (Haberman et al., 2017). Following that, the target RNPs are immunopurified by incubating the lysate with magnetic beads coated with an antibody specific to the RBP of interest. The beads are then subjected to stringent washing conditions using high-salt buffers and ionic detergents to retain only the crosslinked protein-RNA contacts. Next, 3′-ends of RNA fragments immobilised on the beads are dephosphorylated, followed by ligation of an adapter sequence, which creates an annealing site for a primer used in RT. The RNA is then radio-labelled at the 5′-end and eluted from the beads by applying heat (Konig et al., 2011; F. C. Y. Lee & Ule, 2018).

To purify the target RNPs, the eluted protein-RNA complexes are separated on an SDS-PAGE gel, followed by their transfer to a nitrocellulose membrane to separate RBP-bound from free RNAs. When the membrane is visualised as an autoradiograph, the isolated RNPs should be visible as a smear above the combined molecular weight of the RBP and the 3′-adapter (Konig et al., 2011; F. C. Y. Lee & Ule, 2018). This region is then excised from the membrane and used for RNA extraction. In this process, proteins are first digested using Proteinase K, and then the RNA is isolated using phenol/chloroform extraction. Obtained RNA fragments are reverse transcribed into cDNA by adding an RT primer that includes a barcode to allow for multiplexing of samples in sequencing (König et al., 2010; Konig et al., 2011; F. C. Y. Lee & Ule, 2018).

Next, the resulting cDNAs were originally purified from the Urea-PAGE gel to discard RT primers from the mixture (König et al., 2010). In later protocols, this step was replaced by bead-based size selection of cDNAs (Buchbender et al., 2020; F. C. Y. Lee et al., 2021).

After RT, a PCR primer is added to the 3′-ends of cDNAs to generate suitable templates for PCR amplification. In iCLIP, this is achieved by circularising the cDNAs. Then, the cDNAs are amplified by PCR to generate libraries suitable for high-throughput sequencing. The addition of a random sequence into the 5′ PCR primers, i.e., unique molecular identifier (UMI), enables the differentiation between reads resulting from the different RNA fragment after PCR amplification (König et al., 2010).

Since the development of iCLIP, several protocols have been derived to increase its efficiency and reduce experimental time (F. C. Y. Lee & Ule, 2018). In recent years, iCLIP protocols have been enhanced by combining its original steps with those of other truncation-based CLIP methods. These include eCLIP (described below) and irCLIP that replaced radioactive with infrared-dye labelling (Zarnegar et al., 2016). In 2020, iCLIP2 introduced separate adapter ligations, in place of circularisation and linearisation steps, and an additional cDNA amplification step, and bead-based size selection of cDNAs (Buchbender et al., 2020). In 2021, an improved iCLIP protocol (iiCLIP) adopted the irCLIP approach of infrared labelling, while also improving the efficiency of the enzymatic steps and using bead-based size selection of cDNAs (F. C. Y. Lee et al., 2021). Lastly, the iCLIP-1.5 protocol introduced in 2023 incorporated enhanced adapter ligation and the preparation of size-matched input control (SMInput) from the eCLIP procedure and improved cDNA circularisation (Nabeel-Shah & Greenblatt, 2023).

1.1.1.1.2 eCLIP

Like iCLIP, eCLIP is a truncation-based CLIP method that relies on UV-C light crosslinking (Van Nostrand et al., 2016). The two methods differ mainly in their extent of quality control of isolated RNPs, and cDNA library preparation.

After immunoprecipitation, eCLIP omits radiolabelling of RNAs and directly ligates a 3′-adapter to RNA fragments. RNPs are then separated on SDS-PAGE and transferred to a nitrocellulose membrane, from which the relevant RNPs are then isolated. In contrast to iCLIP, which relies on the visualisation of RNPs on the membrane, eCLIP isolates RNPs based on their predicted molecular weight. Therefore, to obtain the RNA-fragments bound by the target RBP, a pre-determined size range of 75 kDa is excised above the predicted molecular weight of the RBP together with the 3′-adapter. eCLIP then follows the same steps as iCLIP to obtain cDNAs, truncated at the crosslink site (F. C. Y. Lee & Ule, 2018; Van Nostrand et al., 2016).

In contrast to iCLIP's circularisation approach for introducing PCR primer binding sites, eCLIP adds a 3′ ssDNA primer to cDNA through an additional ligation step. Ligation conditions with increased concentrations of T4 RNA ligase, polyethylene glycol, and DMSO improve ligation efficiency over iCLIP's circularisation approach and decrease the loss of RNA fragments due to failed ligation (Van Nostrand et al., 2016). These modifications shorten experimental time and produce libraries with more unique RNA fragments, however, they sacrifice some specificity and quality control, which are enhanced by gel visualisation (Hafner et al., 2021; Van Nostrand et al., 2016).

To mitigate the detrimental effects caused by a lack of visualisation, eCLIP produces a size-matched input control experiment (SMInput) to capture non-antigen-specific background. SMInput is produced in parallel to eCLIP by directly transferring 2% of cell lysate containing crosslinked protein-RNA fragments to the gel without performing IP to enrich for the target RBP (Van Nostrand et al., 2016). SMInput is used in computational processing of eCLIP sequencing data to determine RBP binding sites on the transcriptome. A region on the transcriptome is considered a candidate binding site if it exhibits significant enrichment of RNA fragments crosslinked to the target RBP relative to background, represented by the SMInput (Van Nostrand et al., 2016).

1.1.1.1.3 PAR-CLIP

In contrast to iCLIP and eCLIP, PAR-CLIP uses crosslinking at a lower energy UVA/B light (wavelength between 312 and 365 nm) coupled with the use of photoreactive ribonucleoside analogues, such as 4-thiouridine (4SU) and 6-thioguanosine (6SG) (Garzia et al., 2017; Hafner et al., 2010). 16 hours prior to crosslinking, the photoreactive thioribonucleosides are added to living cells and become incorporated into the nascent RNA molecules (Danan et al., 2016). Upon irradiation with UVA/B light, they form crosslinks with bound RBPs. When using 4SU, proximal amino acids form a covalent bond to position 4 of the nitrogen base, thus changing its Watson-Crick base-pairing properties (Ascano et al., 2012). This change leads to a characteristic mutation (T-to-C when using 4SU and G-to-A when using 6SG) when the isolated RNA fragments are reverse transcribed and amplified by PCR, which allows for precise identification of crosslink sites from sequenced data (Garzia et al., 2017; Hafner et al., 2010). Thus, PAR-CLIP does not use read truncations to determine crosslink sites, as is the case in iCLIP and eCLIP, but rather relies on nucleotide transitions, i.e., mutations in the sequenced reads. Due to this set-up, PAR-CLIP employs a different mode of library preparation. While iCLIP and eCLIP ligate only the 3′-adapter to RNA before RT to capture truncated reads, PAR-CLIP already ligates both 3′- and 5′-adapters before RT to serve as binding sites for PCR primers (Danan et al., 2016; Hafner et al., 2010).

Furthermore, PAR-CLIP differs from iCLIP and eCLIP in RNA fragmentation and purification steps. In addition to RNase treatment in the lysate, PAR-CLIP employs on-bead RNA fragmentation and uses different nucleases (RNAse T1 and MNase) with distinct sequence preferences for RNA cleavage (Hafner et al., 2010; F. C. Y. Lee & Ule, 2018). Early versions of PAR-CLIP omitted the membrane transfer step and purified RNA fragments directly from the SDS-PAGE gel (Hafner et al., 2010). However, more recent protocols also perform the nitrocellulose membrane transfer (Garzia et al., 2017).

PAR-CLIP emerged at about the same time as iCLIP and quickly became adopted by several labs (F. C. Y. Lee & Ule, 2018). Due to its design, PAR-CLIP is particularly useful for studying RBP binding to nascent RNAs, and the use of photoactivatable nucleotides increases crosslinking efficiency (Hafner et al., 2021). However, long incubations with photoreactive thioribonucleosides can lead to cellular stress, potentially distorting physiological RBP binding in cells (Huppertz et al., 2014).

1.1.1.1.4 Experimental controls in CLIP

When performing CLIP experiments, it is essential to include appropriate controls to aid in quality control, interpretation, and computational analysis of the data (Haberman et al., 2017; Hafner et al., 2021; F. C. Y. Lee & Ule, 2018; Ule et al., 2003). Common controls for CLIP experiments include the "no-crosslinking" and "no-IP" controls (Hafner et al., 2021). The "no-crosslinking" control omits the application of UV crosslinking to the sample. Upon visualisation on an SDS-PAGE gel, such control should only produce a band corresponding to the molecular weight of the target protein, without any smear resulting from crosslinked RNAs of varying molecular weights. This control verifies the efficiency of crosslinking.

There are two distinct types of "no-IP" controls used in CLIP experiments. The first is performed by simply applying the crosslinked RNPs in the lysate to SDS-PAGE, without enriching for the target RBP. This enables the evaluation of target RBP expression and to obtain a general experimental background of RBP-bound fragments in the sample. This control is predominantly leveraged by eCLIP, i.e., SMInput, to computationally identify regions on the transcriptome where RBP binding is enriched in the samples relative to the general experimental background (Van Nostrand et al., 2016). The other type of "no-IP"

control performs the steps of IP on the sample, however without the antibody against the target RBP. To produce such control, magnetic beads are usually coated with IgG antibody, which is not expected to bind to RNPs in eukaryotes. This type of control allows to assess potential impurities, unrelated to the target RBP, that can be carried over during the sample preparation. Bands observed in this type of "no-IP" control on SDS-PAGE represent non-specific background contaminants transferred from the lysate, such as abundant RNAs present in the sample. The visualisation of the "no-IP" control informs how samples should be excised from the membrane to minimise contributions from non-specific impurities (Hafner et al., 2021; Ule et al., 2003).

Finally, CLIP experiments employing visualisation on SDS-PAGE are typically optimised by preparing low and high RNase controls (Hafner et al., 2021; König et al., 2010). This step is critical for adjusting the level of RNA digestion to an optimal range. To maximise the sensitivity in CLIP experiments, reads must be sufficiently long to enable successful mapping to the genome. In low RNase conditions, a smear should be visible on an SDS-PAGE gel. Conversely, in high RNase conditions, the band on the gel should be centred closer to the combined molecular weight of the RBP and the 3′-adapter.

1.1.1.1.5 Computational analysis of CLIP reads

Reads produced by sequencing reverse-transcribed and PCR-amplified RNA fragments—CLIP reads—are the starting point of computational CLIP data analysis. The standard workflow for the analysis of CLIP reads is composed of four core steps: read pre-processing, mapping of reads to the reference genome, identification of peaks and finally, recognition and characterisation of binding motifs (Chakrabarti et al., 2018; X. Chen et al., 2019; De & Gorospe, 2017).

During pre-processing, reads from multiple samples are first demultiplexed using the barcode sequences in the 3′-adapters. Next, the adapter sequences are trimmed, and PCR duplicates are removed using UMIs. Pre-processed sequences are then aligned to the source organism's genome, and obtained genome coordinates are used to locate the transcript regions bound to the RBP of interest (X. Chen et al., 2019; De & Gorospe, 2017). After the mapping of reads to the genome, the precise locations of crosslink sites can be identified. In truncation-based CLIP methods, crosslinks are determined 1 nucleotide upstream of the 5′-cDNA truncations. In transition-based methods like PAR-CLIP, mapped reads are analysed for characteristic T-to-C or G-to-A mutations.

Crosslink sites or, in rare cases, mapped CLIP reads are then processed with peak-calling approaches to identify clusters, representing potential binding sites of the RBP (Chakrabarti et al., 2018; De & Gorospe, 2017). For this purpose, a wide range of peak-calling tools are available, which employ distinct approaches and algorithms to identify peaks of CLIP signal (Bischler, 2017; Boyle et al., 2023; Capitanchik et al., 2022; Curk, 2019; Krakau et al., 2017; Schwarzl et al., 2022; A. Shah et al., 2017; Uren et al., 2012). The variety of peak-callers and their adjustable parameters address diverse needs. Some are better suited to finding a small number of highly confident binding sites, and others to finding many potential binding sites. Ideally, the peak caller would minimise the number of false-positive and false-negative sites, yielding high specificity sites while retaining as many relevant binding sites as possible. The biological relevance of identified peaks can be ascertained from the proximity of RBP-binding motifs and the enrichment of peaks around relevant regulatory sites, such as, for example, intron-exon boundaries in the case of splicing factors or 3′-UTR termini for factors that control 3′-end mRNA processing (Chakrabarti et al., 2018; Haberman et al., 2017). For our analyses we used the Clippy peak-caller (see Methods, (Kuret et al., 2022), as its algorithm accounts for important aspects of RNA

biology, like splicing and variability in expression levels, while offering high speed and a wide array of parameter options to fine-tune the peaks.

Finally, the resulting peaks can be used to determine RBP binding specificity, including the recognition of specific types of nucleotide sequence motifs, their relative positioning and density, RNA structure, and RNA-modifications (Chakrabarti et al., 2018; X. Chen et al., 2019). Moreover, they can be analysed in the context of important regulatory elements to indicate the involvement of an RBP into processes related to RNA metabolism; for the presence of genomic variants and their phenotypical impact (Romo et al., 2023); or in the context of bound RNA targets—using gene ontology—to provide insight into the cellular pathways that are regulated by the studied RBP.

1.1.1.2 *In vitro* approaches for evaluating the specificity of RNA-binding proteins

In vitro approaches for evaluating the specificity of RNA-binding proteins aim to identify the features that drive their RNA recognition. These methods involve the incubation of the purified recombinant RBP with a pool of RNA sequences and then enrich for the RBP-bound RNAs (Dasti et al., 2020). Examples of such methods include RNA-Bind-n-Seq (RBNS) (Lambert et al., 2014), RNAcompete (Ray et al., 2009) and Exponential Enrichment Selection (SELEX) (Ellington & Szostak, 1990; Jolma et al., 2020; Tuerk & Gold, 1990). Typically, these methods measure the affinity of the full RBP or its specific RNA-binding domain towards shorter RNA sequences (~40nt), which are obtained either by transcription of random DNA fragments, i.e., natural sequences (Jolma et al., 2020; Lambert et al., 2014), or by synthetically creating a pool of all possible RNA sequences of certain length (Lambert et al., 2014; Ray et al., 2009). Besides the generation of the RNA pool, the methods also diverge in their RNA enrichment step. RNAcompete and RBNS perform a single binding cycle, while SELEX performs multiple incubations, starting with the initial RNA pool and using enriched sequences, amplified with PCR, in the subsequent binding cycle (Dasti et al., 2020). RBNS also has a unique advantage of being able to detect high- and low-affinity sequences, because the target protein is incubated with the RNA at different concentrations (Lambert et al., 2014). However, large amounts of protein are required for such an experiment, which can be challenging in some cases, as RBPs often contain intrinsically disordered regions (IDRs) that are hard to purify (Dasti et al., 2020). After the enrichment of bound RNAs, the bound fragments are sequenced and used to characterise the features recognised by the RBPs. In vitro methods can produce information of RBP binding preferences in the form of linear sequence motifs, as well as structural motifs, given the RNAs in the pool are sufficiently long (Jolma et al., 2020; Ray et al., 2009).

Thus, in vitro methods are extremely valuable, as they can provide information of intrinsic RBP specificity, at a level of the whole protein, or its individual RBDs. The use of *in vitro* approaches orthogonally to CLIP allows to disentangle the RBP-specific binding sites identified *in vivo* from those that arise because of experimental artefacts, co-purified RBPs, or general sequence preferences of RBP-bound transcript regions (Hafner et al., 2021). In this dissertation, I will systematically compare sequence preferences identified for the RBPs *in vitro*, with those identified from CLIP data, to elucidate the influence of experimental biases in CLIP on enriched motifs, assess the quality of CLIP data and benchmark the accuracy of motif-discovery tools.

1.1.1.3 Motif discovery

Many computational tools exist to discover overrepresented motifs in any set of unaligned biological sequences. The use and development of motif discovery algorithms surged with methodological advancements in the field of transcription factor research, to identify

regulatory motifs recognised by transcription factors on DNA. Therefore, initial tools for motif discovery focused on linear sequence motifs, however, in contrast to DNA, RNA molecules arrange into diverse structural conformations that can contribute to binding of RBPs (Dominguez et al., 2018; X. Li et al., 2010; Ray et al., 2013). Today, various tools exist for motif discovery from RNA molecules, but methods such as CLIP present unique challenges due to technical biases such as preferential crosslinking to certain nucleotides (Knörlein et al., 2022), sequence preferences of RNA fragmentation, and influence of abundant transcripts or repetitive elements (Hafner et al., 2021; Pietrosanto et al., 2021). In the following section, we review the different types of tools used for motif discovery from RNA and highlight their strengths and limitations.

1.1.1.3.1 Motif-discovery tools for characterisation of RBP binding motifs

Motif discovery is a challenging problem because it involves finding short, similar sequences (needles) in much longer sequences (haystacks) (Bailey et al., 2006). To tackle this, motif finding tools employ two different approaches to modelling foreground and background: signal-only learning and discriminatory motif discovery (Maaskola & Rajewsky, 2014). In signal-only learning, the foreground sequences and background sequences are modelled within the same input sequence – the RBP-bound motifs represent foreground and the rest of the sequence represents background. Conversely, in discriminatory motif discovery, the foreground and background are modelled with different sets of input sequences (Maaskola & Rajewsky, 2014).

Existing tools employ diverse strategies to motif discovery. They encode input sequences with varying levels of complexity, represent discovered motifs in various ways, and learn the motifs using a combination of various algorithms, such as k-mer enrichment, expectation-maximisation, Hidden Markov Models (HMMs), and deep learning (Pietrosanto et al., 2021; Sasse et al., 2018). Based on input encoding, the tools can be broadly stratified into two classes: 1) the tools that only encode input with the primary nucleotide sequence (Agostini et al., 2014; Bailey et al., 2006; Leibovich et al., 2013), 2) tools that incorporate secondary RNA structure into the model. Furthermore, the structure-based models can be separated into those that encode structure at a level of a single nucleotide in a linear fashion (Bahrami-Samani et al., 2015; Budach & Marsico, 2018; Hiller et al., 2006; Kazan et al., 2010; Orenstein et al., 2016), and tools that use more complex encoding of nucleotide sequence and RNA structure together, preserving base-pairing information (Maticzka et al., 2014).

1.1.1.3.1.1 Sequence-based models

Sequence-based models encode the input RNA set only with their primary nucleotide sequence, using the alphabet to represent the four nucleotides (A for adenosine, C for cytidine, G for guanosine and T or U for thymine or uridine, respectively). Most sequence-based tools perform *de novo* motif discovery, generally by combining k-mer modelling with expectation-maximisation and/or HMMs to find common patterns in the representative set of sequences (Dasti et al., 2020; Sasse et al., 2018). Most of the early tools for motif discovery, for example MEME and GibbsSampler, model only the positive sequences, i.e., perform signal-only learning (Maaskola & Rajewsky, 2014). Signal-only learning was particularly beneficial prior to the wide-spread accessibility of high throughput technologies. Such approaches allow motif finding even in small sets of sequences, as they represent both background and foreground within the same input sequence – the foreground corresponds to the binding motif, and the background does not. Most sequence-based tools use probabilistic models to describe the motifs and encode them with one or several Position Weight Matrices (PWMs) (Dasti et al., 2020; Sasse et al., 2018). PWM indicates

the probability for each position of the motif to incorporate a given nucleotide (Stormo et al., 1982). Nevertheless, sequence-based models differ in their underlying assumptions related to the input sequences and binding motifs. For example, MEME, DRIMust, and SeAMotE assume that nucleotides in the motif are independent of each other and that the motifs are ungapped, i.e. the models do not allow motifs with insertions or deletions (Dasti et al., 2020). Furthermore, expectation maximisation models parametrise motif length, a number of motifs per sequence, and look for motif positions within a sequence (i.e. offset) (Bailey et al., 2006; Dasti et al., 2020). In contrast, k-mer-based approaches are more flexible and allow for varying lengths of motifs, while also preserving higher-order dependencies between the nucleotides (Sasse et al., 2018). Due to their underlying assumptions, expectation maximisation models provide accurate results for RBPs that bind strongly to a short and specific uni-partite motifs. Conversely, they are not well suited for RBPs that bind bi- or tri-partite motifs, repetitive RNA sequences and degenerate motifs (Dasti et al., 2020). Despite these limitations, tools based on expectation maximisation remain among the most widely used approaches for motif discovery from RNA, due to their accessibility and ease of use.

To address some of these limitations, more complex tools such as NRLB and GLAM2 model the problem as expectation-maximisation calculation in conjunction with HMM, while also considering potential dependencies between nucleotides and allowing for gaps in the discovered motifs, making them suitable for RBPs that bind complex patterns (Frith et al., 2008; Rastogi et al., 2018).

1.1.1.3.1.2 Sequence and structure-based models

Adding structural context to linear sequence for motif discovery from RNA molecules has highlighted the importance of the RNA structure in motif recognition by RBPs (X. Li et al., 2010; Ray et al., 2009). As a result, several motif discovery tools have been developed that consider the structural context in addition to the primary sequence when identifying RBP binding sites (Dasti et al., 2020; Pietrosanto et al., 2021; Sasse et al., 2018). These tools employ external software, such as RNAfold (Hofacker, 2003; Lorenz et al., 2011) and Sfold (Ding et al., 2004), that predict RNA secondary structure from the input nucleotide sequence in a way which minimises its predicted free energy (X. Wang et al., 2023). The representation of structure and its use varies between motif finders (Sasse et al., 2018).

Some tools, such as MEMERIS, use structure only as an exclusion criterion to focus the finding of linear sequence motifs toward unstructured regions (Hiller et al., 2006). This approach is based on the idea that sequence motifs are only accessible to RBPs in unstructured regions of RNA. Other tools, such as RNAcontext, RCK, and PRIESSTESS, use structure proactively to identify specific "structural motifs" in addition to sequence determinants recognised by RBPs (Budach & Marsico, 2018; Kazan et al., 2010; Laverty et al., 2022; Orenstein et al., 2016). These models encode predicted secondary structure for each input RNA sequence on a nucleotide-level, using a structural alphabet, for example denoting hairpin loops with H, stems with S, etc. (Sasse et al., 2018). More complex models, such as GraphProt, encode RNA structure and nucleotide sequence together, to preserve information of base pairing for exact nucleotides within sequence (Maticzka et al., 2014). Due to their complex representations of the input data, these methods can capture the subtle dependencies between sequence and structure in the motifs. However, this complexity makes them more difficult to interpret (Laverty et al., 2022; Sasse et al., 2018).

Despite their differences, all structure-based motif finders are limited in their capacity to model RNA structure. These models can only efficiently predict short-range intramolecular secondary structures, however in biological systems structural motifs may arise on long-range structures (hundreds of nucleotides apart), tertiary structures and from

intermolecular RNA-RNA contacts. Such global RNA structure cannot be captured by modelling input sequences. In the future, we expect to see an increased use of structure sequencing in conjunction with nucleotide sequence to represent cellular RNA-structure in motif discovery (Morandi et al., 2022). Experimental techniques that employ sequencing to probe RNA structure transcriptome-wide have recently emerged (Marinus et al., 2021). The readouts of such experiments can be used in conjunction with CLIP or in vitro experiments to precisely determine how structural elements contribute to RBP binding. Providing experimentally characterised structure is more accurate than using structure predictions, based on sequence, as predictions fail to account for long-range intramolecular and intermolecular RNA structures present in cells.

1.1.1.3.2 Motif discovery from CLIP experiments

When conducting discriminatory motif discovery from in vitro experiments, full-length sequences of RBP-bound RNAs are used as foreground sequences, while the background is represented with a random pool of RNAs used to conduct the experiment (Kazan et al., 2010; Laverty et al., 2022). In contrast, for determination of RBP binding motifs from CLIP-seq data, foreground sequences are extracted from genomic regions with a high density of CLIP signal, i.e., peaks. To construct background sequences for CLIP-seq, common practices include reshuffling the foreground sequences, or selecting relevant background sequences from the dataset. Background sequences are not expected to be specifically enriched in target RBP binding; therefore, they can be obtained from fixed regions around the foreground sequences (Feng et al., 2019), or by randomly sampling regions from the same gene as the foreground sequences matched for length and transcript region (Van Nostrand et al., 2020).

With advancements in CLIP technologies, motif discovery algorithms have adapted to leverage the high resolution of these methods. Including the information on crosslink positions into the model increases the accuracy of motif discovery, as crosslinks tend to coincide or be in proximity of the motifs that are recognised by the RBP (Bahrami-Samani et al., 2015; Feng et al., 2019). Currently, two tools implement the strategy of joint modelling of RNA sequence and crosslink positions: Zagros (Bahrami-Samani et al., 2015) and mCross (Feng et al., 2019). Both tools employ a probabilistic model to derive motifs but use information of crosslinking positions in different ways. The model used by Zagros assigns greater weight to motifs that are in proximity of crosslinks—the distance between the motifs and crosslink sites is factored into probability calculations. mCross, however, first uses crosslink sites to identify genomic regions with high density of crosslink signal and uses these to find short, enriched motifs (k-mers). These k-mers then serve as seeds to initialise the probabilistic model which performs the motif discovery task, by jointly modelling nucleotide sequence and precise positions of protein-RNA crosslink sites, as opposed to distance from the crosslink sites (Feng et al., 2019).

However, motifs enriched at crosslink sites can also reflect technical biases of CLIP experiments, such as preferential crosslinking to uridines, sequence preferences of RNA fragmentation or ligation biases. Motif discovery from CLIP datasets is also impeded by the imbalanced nucleotide composition of different genomic regions, the presence of repetitive elements and the differences in transcript expression (Hafner et al., 2021). Nevertheless, these challenges can be mitigated by thoughtfully modelling background sequences to capture these biases without exhibiting specific enrichment of RBP binding. While existing tools have leveraged crosslink sites to improve the selection of foreground sequences, they have not incorporated approaches that effectively mitigate the biases introduced by CLIP experiments.

1.1.1.3.3 Implicit motif discovery with deep learning models

The trend in motif discovery has significantly shifted in recent years, transitioning from explicit modelling to implicit learning through deep learning models. This change is exemplified by models such as DeepBind, pysster, iDeep, DeepRiPe, and RBPNet (Alipanahi et al., 2015; Budach & Marsico, 2018; Ghanbari & Ohler, 2020; Horlacher et al., 2023; Pan & Shen, 2017).

Pysster, for instance, uses a convolutional neural network to categorise nucleotide sequences into user-defined classes (Budach & Marsico, 2018). Through this process, it implicitly learns the nucleotide and structural motifs associated with each class; this can be leveraged to learn RBP binding motifs from CLIP data, by training the model to differentiate between sequences within the CLIP peaks, i.e., RBP-bound, and not in CLIP peaks, i.e., unbound. If the peaks represent binding sites that are specific to target RBP, the algorithm will learn the RBP-binding motifs, associated with these sequences. In contrast, iDeep is a generative model, which learns the RBP binding motifs by training to predict the nucleotide sequences of RBP-bound RNA molecules enriched in *in vitro* RNAcompete experiment (Pan & Shen, 2017). DeepBind is a convolutional neural network model, which predicts an enrichment score of a sequence in the RNAcompete assay, thereby learning the features that mediate the RBP binding to that sequence (Alipanahi et al., 2015). Finally, models like DeepRiPe and RBPNet are designed to predict the locations of RBP binding sites on transcripts (Ghanbari & Ohler, 2020; Horlacher et al., 2023). RBPNet predicts the distribution of CLIP signal while DeepRiPe predicts peak regions on the RNA. Both models train on RNA sequence and learn the motifs associated with peaks of crosslinking signal in CLIP. These motifs can then be extracted from the model's predictions using feature-importance analyses.

Deep learning models are widely applicable in various biological contexts, as they make fewer assumptions about the input data and can encode complex relationships between different motifs. However, they also present challenges in terms of interpretation, accessibility, and implementation for individual researchers. For instance, the association of a motif with CLIP peaks does not necessarily imply that the motif is directly bound by the target RBP, which is a common issue also when applying traditional motif-finding approaches to CLIP data. To enhance the accuracy of implicit motif discovery, the training input data must be highly specific and of high quality, necessitating controlled experiments with a low level of background noise. Furthermore, these models need to better distinguish true biological signals from technical noise. A recent model, RBPNet, aims to address this issue by simultaneously modelling CLIP signal and SMInput control signal, which represents non-antigen-specific background (Horlacher et al., 2023). It is also worth noting that implicit learning of motifs tends to take longer than explicit modelling because the model needs to be trained first, then the prediction is made on test data, followed by a feature-importance analysis to extract the learned motifs. Despite these challenges, these models will play a central role in the future of motif discovery.

1.1.2 Resources of protein-RNA interaction data

In recent years, the number of available CLIP datasets has rapidly increased, which opened the potential for systematic analyses of the features that recruit RBPs to specific RNA sites. Raw CLIP-seq reads can be obtained for individual publications from the central repositories, such as the Gene Expression Omnibus (GEO) and the Sequence Read Archive (SRA). However, collection and analysis of these fragmented datasets is challenging, due to the lack of uniform processing pipelines, and batch variation between different labs. Therefore, curated collections of CLIP data are particularly valuable. CLIP peaks from various datasets are integrated in the POSTAR database, where they can be explored in

the context of functional annotations, genetic variation, disease mutations, RNA structure, translation and miRNA-mediated RNA degradation (Zhao et al., 2022). To date, the largest resource of CLIP data was produced with eCLIP method and published by the Encyclopaedia of DNA Elements (ENCODE) consortium (ENCODE Project Consortium, 2012; Van Nostrand et al., 2016). In addition to this, a wealth of PAR-CLIP data was produced by publications (Hafner et al., 2010; A. S. Y. Lee et al., 2015; G. Martin et al., 2012). Currently, the datasets available on ENCODE and on POSTAR are the most widely used for benchmarking computational models geared towards characterising RBP binding or predicting RBP binding from CLIP.

Besides CLIP data, two large collections of *in vitro* protein-RNA interaction data were published. The first was produced with the RNAcompete method and encompassed in total 80 distinct RBPs, while the second was produced with RBNS and encompassed 78 RBPs (Dominguez et al., 2018; Ray et al., 2013). The results from the RBNS assay were also integrated into ENCODE (Van Nostrand et al., 2020).

In this work, we utilise all the aforementioned datasets (eCLIP, PAR-CLIP, RNAcompete and RBNS) to benchmark the specificity of motif discovery with PEKA from CLIP data, and to study how inherent experimental biases of CLIP affect enriched motifs. Since the publication of this work, another large RNAcompete dataset was produced for 492 RBPs that lack canonical RBDs (Ray et al., 2023). Furthermore, a web platform Flow has been developed to facilitate standardised, quality-controlled, reproducible, and traceable analysis of CLIP and other high-throughput sequencing data. Flow's vision extends beyond data analysis; it aspires to construct a database of curated CLIP datasets that will be accessible to researchers worldwide (Capitanchik et al., 2023).

1.1.3 LIN28A and its roles in early embryonic development

LIN28A is an essential RBP that plays diverse roles in cellular processes (Tsialikas & Romer-Seibert, 2015; Wu et al., 2022). It consists of two distinct RBDs that provide unique modes of RNA recognition: a pair of zinc-finger domains (ZnF) binds to GAGG motifs within hairpin loops, while the Cold-Shock domain (CSD), structured as a beta-barrel, binds to GAU consensus motifs (Ustianenko et al., 2018; Wilbert et al., 2012; Yamamoto et al., 2022). LIN28A, along with its homolog LIN28B, is the only animal protein with this combination of domains, which enables its diverse regulatory functions in cells (Tsialikas & Romer-Seibert, 2015).

LIN28A is highly expressed during early embryonic development; however, its expression decreases as cells differentiate. In a developing organism, the expression of LIN28A is constrained to specific tissues and by adulthood, the protein remains expressed only in a few epithelial tissue layers in the kidney, muscle, and erythrocytes (de Vasconcellos et al., 2014; D. H. Yang & Moss, 2003). LIN28A regulates the timing of embryonal progression, allowing for normal embryo development and growth-related and tissue-repair processes later in life. It controls these processes by interacting and acting on microRNAs (miRNAs), mRNAs, and even DNA to modulate protein expression, splicing, and transcription (Wu et al., 2022).

1.1.3.1 Versatile regulatory roles of LIN28A in physiology and disease

LIN28 was discovered in C. elegans in 1984 as a heterochronic gene that regulates developmental timing (Ambros & Horvitz, 1984). This function of LIN28 has since been observed in Drosophila, Xenopus, zebrafish and mammals (Tsialikas & Romer-Seibert, 2015). LIN28A promotes the proliferation of embryonic and cancer stem cells and increases the reprogramming of somatic into pluripotent cells when introduced with other

reprogramming factors (Hanna et al., 2009; Shyh-Chang & Daley, 2013; J. Yu et al., 2007). On the organismal level, LIN28A promotes growth by increasing insulin sensitivity, glucose uptake and controlling a shift from oxidative metabolism towards glycolysis (Docherty et al., 2016; Ma et al., 2014; H. Zhu et al., 2010). Thus, LIN28A is an oncogene, and its dysregulation in tissues can lead to various types of aggressive cancers (Balachandran & Narendran, 2023; H. Wang et al., 2016; Wu et al., 2022). Conversely, reduced LIN28A expression can lead to dwarfism and other growth defects (Wu et al., 2022).

The control of developmental timing by LIN28 was predominantly studied in the context of its regulation of the let-7 pathway. let-7 are a conserved class of miRNAs, which negatively regulate cell proliferation and differentiation by inhibiting the expression of proteins necessary to maintain the proliferative properties of stem cells (Roush & Slack, 2008). LIN28A and LIN28B homologs cooperatively inhibit the biogenesis of let-7 miRNAs, thus maintaining the proliferative properties of naïve embryonic stem cells (Piskounova et al., 2011). LIN28A and LIN28B inhibit let-7 biogenesis through two distinct mechanisms. LIN28B binds to let-7 precursors in the cell nucleus and prevents their export into the cytoplasm, where they are processed by Dicer enzyme. It also prevents their nuclear processing by Drosha enzyme (Piskounova et al., 2011). In contrast, LIN28A binds let-7 precursors in the cytoplasm and mediates their uridylation by recruiting terminal uridyl transferases. The uridylation of the let-7 precursors blocks their processing to mature let-7 by Dicer enzyme and leads to their degradation (Heo et al., 2008, 2009).

Because of its role in let-7 pathway, the differentiation of cells is linked to LIN28A repression in most mammalian models. However, two studies found that LIN28A is also required for efficient differentiation to certain cell types, independently of let-7 (Faas et al., 2013; Polesskaya et al., 2007). Both studies observed promoting effects of LIN28A on mRNA translation. Since then, a wide range of studies reported diverging effects of LIN28A on the mRNA; the majority reported promoting effects on translation, while one reported promoting and repressive effects (Cho et al., 2012; Peng et al., 2011; Polesskaya et al., 2007; Shyh-Chang et al., 2013; Wilbert et al., 2012; Zhang et al., 2016). Recently, other let-7 independent effects of LIN28A were observed: In the context of breast cancer, LIN28A controls the decay of specific mRNAs to promote tumorigenesis (Zou et al., 2022); let-7 independent pathways of LIN28A are also critical in tissue repair and regulate cell metabolism and migration (Shyh-Chang et al., 2013). Beyond this, LIN28A is linked to brain neurogenesis (Hu et al., 2022), Parkinson's disease (Chang et al., 2019), cellular senescence (Broughton et al., 2022), and the growth of skeletal muscle (Polesskaya et al., 2007). The mechanisms invoked in these processes are still largely unknown but could be linked to the reshaping of RNA structure by recognition of RNA modification (Sun et al., 2019), localisation of LIN28A and its partner RNAs to cellular condensates, such as P-bodies and stress granules (Balzer & Moss, 2007), translational effects of LIN28A, as well as mRNA decay tumorigenesis (Zou et al., 2022).

1.1.3.2 LIN28A in early embryonic development

A study by Tsanov et al. showed that during early embryonic development—when naïve embryonic stem cells transition to primed pluripotency—LIN28A is phosphorylated by the MEK/ERK signalling pathway, and that this phosphorylation promotes naïve-to-primed transition (Tsanov et al., 2017). However, the exact mechanism through which this phosphorylation of LIN28A promotes naïve-to-primed transition remains unclear (Tsanov et al., 2017).

The cell-fate transition from naïve to primed pluripotency is precisely coordinated by the pivot from WNT to ERK signalling that occurs at the rosette stage, during the first 24h after blastocyst implantation in mice development (Neagu et al., 2020). This shift in

cell-signalling dynamics is accompanied by the decline in transcription factors that maintain the naïve pluripotency expression programme, such as Nanog, Esrrb, and Klf proteins (M. Li & Izpisua Belmonte, 2018). Previous work from our lab showed that LIN28A regulates the clearance of these naïve-pluripotency factors by mediating their mRNA decay (Modic et al., 2021). Modic et al. showed that transcripts undergoing selective decay in naïve-to-primed pluripotency transition exhibit higher levels of LIN28A-binding in their 3′-UTR regions compared to other genes. Moreover, the binding of LIN28A on these transcripts further increases after the activation of MEK/ERK signalling. This agrees with findings from Tsanov et al., who showed that phosphorylation stabilises LIN28A and thus increases its cellular content, as well as promotes the interactions with mRNAs (Tsanov et al., 2017). In contrast to Modic et al., Tsanov et al. proposed a role of pLIN28A in mRNA translation.

In mice, the phosphorylation site of LIN28A is located at a residue S200 in its C-terminal intrinsically disordered region. Despite intrinsically disordered regions being associated with low evolutionary conservation, this phospho-site is evolutionarily conserved (Tsanov et al., 2017)—suggesting an important role in controlling LIN28A function. In this work, we explore the mechanism by which phosphorylation of LIN28A at S200 regulates selective decay of mRNAs encoding for naïve pluripotency factors, and uncover mRNA features that confer transcript susceptibility to pLIN28A-mediated decay.

1.2 Purpose

The purpose of this doctoral dissertation is to develop a versatile computational tool for motif-discovery from CLIP data and apply it in concert with other bioinformatic approaches to interrogate the RNA motifs that specify RBP binding from two different perspectives. In the first part of the dissertation, we use this tool from a systems perspective to study a large resource of CLIP data to examine the impacts of CLIP-related technical biases on detected motifs and the relationships between motif-enrichment patterns and biological properties of RBPs, such as protein structure and binding preferences for specific RNA regions. In the second part of the dissertation, we study how IDRs can modulate RBP specificity from a mechanistic perspective; specifically, we investigate how the phosphorylation of IDR in LIN28A changes its RNA specificity in the context of a dynamic cell-fate transition in early embryonic development. With these complementary perspectives, we aim to establish the framework for the use of motif analysis to study the general properties that govern RBP-binding, to assess the quality of CLIP datasets, and to study dynamic changes in RBP binding preferences in response to cellular signals.

1.3 Goals

The goals of this thesis are aligned with its purposes described in the previous section. The first goal of this dissertation is to develop a computational tool for reliable discovery of RBP binding motifs from CLIP data that effectively reduces the influence of inherent experimental biases on motif discovery. To assess its performance on the task of motif discovery and minimisation of technical bias, the developed tool will be benchmarked against a competing state-of-the-art tool and against motifs discovered from orthogonal *in vitro* datasets that are not subject to the same technical influences as motifs discovered from CLIP.

The second goal of this dissertation is to conduct a large-scale analysis of binding motifs for a compendium of CLIP datasets, to explore how different CLIP methods and properties of different RBPs affect detected binding specificity. Using clustering, we will explore the

relationships between RBP binding specificity and their structural properties, specifically the content of canonical RNA binding domains, IDRs and low-complexity regions, their binding preferences for specific genomic regions, and their detectability by CLIP. We will examine the existence of common motif enrichment patterns that are unique to specific CLIP methods, characterise them, and assess whether the patterns of k-mer enrichment can serve as a signature of data quality.

The third goal of this dissertation is to study the mechanism by which the phosphorylation of IDR in LIN28A mediates selective mRNA decay of naïve-pluripotency factors in early embryonic development. For this, we will use iCLIP data of LIN28A and cytoplasmic poly(A)-binding proteins (PABPs), which are known effectors of mRNA stability. We will examine iCLIP data obtained at different stages of naïve-to-primed transition and employ motif analysis and complementary bioinformatics approaches to study RBP binding on transcripts affected by selective mRNA decay. Specifically, we will investigate RBP binding motifs and the positioning of its binding sites on the transcripts relative to binding sites of PABP proteins that were also implicated in this process. Finally, we will analyse RNA features to identify regulatory elements that predict the susceptibility of transcripts to LIN28A-mediated decay. Through this comparative study of CLIP data in different stages of early embryonic development, we aim to gain mechanistic insights into the role of IDR phosphorylation as an inducible modulator of LIN28A RNA-binding specificity, function, and assembly.

1.4 Hypotheses

When performing motif discovery from CLIP data, the inclusion of crosslink-associated features into the model increases the accuracy of tools that identify enriched motifs (Bahrami-Samani et al., 2015; Feng et al., 2019). However, motifs enriched at crosslink sites can also reflect technical biases of CLIP experiments. We hypothesise that:

1. We can increase the probability of identifying biologically relevant enriched motifs from CLIP data by employing low-count crosslink sites to model background sequences.

2. Datasets produced by each variant of CLIP method will exhibit some common motif enrichment patterns that can inform on its technical biases.

3. The binding specificity of RBPs is affected by the protein's structural features, its IDRs, and its genomic regional binding preferences.

LIN28A is a key effector of selective mRNA decay that drives naïve-to-primed cell-fate transition in early embryonic development. Previous work established that in this transition, LIN28A is phosphorylated in its intrinsically disordered region by the MEK/ERK signalling pathway, which promotes naïve-to-primed transition (Tsanov et al., 2017). Another study showed that during this cell-fate transition, LIN28A is required for selective clearance of naïve regulon and coordination of pluripotency progression (Modic et al., 2021). Tsanov et al. established that phosphorylation of LIN28A promotes its interactions with mRNA partners, and Modic et al. showed that transcripts prone to developmental decay—including naïve regulon—exhibit higher levels of LIN28A binding (Modic et al., 2021; Tsanov et al., 2017). Moreover, PABPs are known effectors of mRNA stability, and in addition to LIN28A, the increased binding of PABPC1 to the mRNA was also predictive for transcript decay (Modic et al., 2021). Previous studies showed that PABPs can interact with LIN28A (Balzer & Moss, 2007) independently of RNA (N.-K. Yu et al., 2021). Based on these previous studies, we hypothesise that:

1. Phosphorylation of LIN28A in its IDR changes its RNA binding properties to activate its function in selective mRNA decay.

2. The transcripts that undergo selective decay exhibit distinctive sequence features that are selectively recognised by phosphorylated LIN28A.

3. Phosphorylated LIN28A acts on the transcripts in concert with cytoplasmic PABPs to regulate selective mRNA decay in naïve-to-primed transition.

1.5 Structure of the Thesis

The remainder of this dissertation is organised into five chapters, as follows. Chapters 2, 3, and 4 showcase the results underpinning the key contributions and represent the main body of the work. Chapter 5 is a discussion, where we summarise the key findings and the limitations of the study, as well as present potential future directions. Lastly, Chapter 6 presents the methodology used to generate the results in this work.

Chapter 2 introduces PEKA, a computational tool developed for motif analysis from CLIP data. This chapter also compares the performance of PEKA with the current state-of-the-art tool and evaluates it against in vitro derived motifs.

Chapter 3 presents a comparative motif analysis across a diverse range of CLIP datasets. This analysis is integrated with various types of orthogonal data to derive insights into RBP sequence specificity and the technical biases inherent in CLIP.

Chapter 4 focuses on the functional analysis of LIN28A during the naïve-to-primed transition in early embryonic development, presenting our key findings in this area.

Chapter 5 discusses the contributions of this work, its limitations, and potential future applications. This chapter provides a comprehensive overview of the impact and implications of our research.

Finally, **Chapter 6** details the methodologies used to produce the results presented in Chapters 2, 3, and 4. This chapter provides a thorough explanation of our approaches, ensuring transparency and facilitating reproducibility of our research.

Chapter 2

Positionally Enriched K-Mer Analysis

The characterisation of RBP binding *in vivo* can provide key insights into the regulatory mechanisms of RBPs in a particular cellular context. Even though CLIP data offers precise positional information of RBP binding on the transcript, which is lacking in *in vitro* methods, only two tools—Zagros and mCross—exist that exploit this advantage for motif discovery (Bahrami-Samani et al., 2015; Feng et al., 2019). Both tools utilise crosslink sites to enhance motif discovery with different levels of precision. Zagros incorporates the distance between sequence motifs and crosslink sites into its probabilistic model, resulting in a greater emphasis on motifs that are near the crosslink sites. mCross, however, identifies enriched k-mers from regions with a high density of crosslinks, which act as seeds for *de novo* motif discovery. Additionally, mCross fine-tunes a likelihood function by precisely registering the positions of crosslinks with respect to motifs. Unfortunately, the advanced mCross model is not available for public use, which restricts its adoption by the research community.

Here, we present positionally-enriched k-mer analysis (PEKA), a computational tool for analysis of linear sequence motifs recognised by RBPs *in vivo*. PEKA uses CLIP data as an input and is designed to reduce the impacts of technical biases and sequence biases of genomic regions on enriched motifs, by modelling low-count crosslinks from the analysed dataset as background and analysing specific types of genomic regions separately. PEKA presents clusters of enriched k-mers, which together with the ability to evaluate distinct genomic regions and handle repetitive elements, makes it an accurate and versatile tool that offers an intricate view into RBP binding specificity. Furthermore, k-mer enrichments reported by PEKA allow for easy quantitative comparisons between different datasets, which is essential for system-level investigation of binding specificity. We show how these features and its accessibility make PEKA a significant improvement over the state-of-the-art method.

PEKA is a practical k-mer enrichment analysis, which leverages precise positional information of crosslink sites while simultaneously minimises crosslink-associated biases to find biologically relevant RBP-binding motifs (Kuret et al., 2022). The following sections are adapted from our publication "Positional motif analysis reveals the extent of specificity of protein-RNA interactions observed by CLIP" (Kuret et al., 2022) and describe: the core steps of PEKA and the reasoning behind their implementation, the capability of PEKA to detect multiple binding modes of RBPs, and its performance compared to state-of-the-art mCross method.

2.1 An Overview of PEKA's Methodology and Logic

PEKA includes several features that are designed to examine and minimise the impact of technical biases of CLIP, and thereby to obtain enriched motifs that mediate RBP binding specificity. PEKA can perform motif discovery either across the full transcriptome or within defined transcriptomic regions (Figure 2.1A), with the provided options including introns, 3′-UTR, remaining exonic regions of protein-coding genes (coding sequence (CDS) combined with 5′-UTR), non-coding RNAs (ncRNAs), and the rest of non-annotated intergenic regions. PEKA also provides an option to include or exclude repetitive regions in the analysis.

Importantly, PEKA implements an approach of background normalisation that aims to minimise the technical biases at crosslink sites. Crosslink sites are determined by the first nucleotide of aligned sequencing reads, which can include nucleotide preferences of UV crosslinking or sequence biases of cDNA ligation (Haberman et al., 2017; Hafner et al., 2021). To normalise for these biases, PEKA extracts the background sequences that are centered on low-scoring crosslink sites (out-of-peak crosslinks, oXn), which are located outside the peaks (i.e., areas with high crosslink density). Conversely, the foreground sequences are centered on high-scoring crosslink sites located inside the crosslinking peaks (thresholded crosslinks, tXn) (Figure 2.1B). The peak regions are provided by the user, and peaks can be identified by any peak calling tool. For the purposes of this study, we used Clippy v1.5.0 (Capitanchik et al., 2022) (Methods).The first step of PEKA is thus to split crosslink sites into thresholded (tXn) and out-of-peak (oXn) sites based on whether the sites overlap with a peak and whether the cDNA count at the site meets a minimum threshold that is defined in a region-specific manner. Notably, tXn and oXn sites are expected to be affected by the same technical biases since they are part of the same cDNA library.

PEKA derives background and foreground sequences from fixed-length genomic windows centered on tXn and oXn, respectively (Figure 2.1C). In the current study, the sequences for motif discovery were defined by expanding 20nt up- and downstream of crosslink sites (total sequence length 41nt); however, window size can be adjusted to search for motifs closer or further away from the crosslink sites. Next, PEKA scans for the presence of k-mers across collected sequences and obtains k-mer counts at each sequence position. K-mer counts are then normalised with the number of evaluated sequences to obtain k-mer occurrences (Figure 2.1D). Afterwards, the relevant positions around crosslink sites are defined at which the enrichment is calculated for each k-mer (Figure 2.1E). For this purpose, relative k-mer occurrence is calculated by normalising the nucleotide-level k-mer occurrences with the mean k-mer occurrence in a distal region (defined as −150…−100 and 100…150 nt around tXn). Relevant positions are those where relative k-mer occurrence of each specific k-mer around tXn is higher than a threshold value that is determined based on relative occurrences of all k-mers (Methods, Figure 2.1E). The relevant positions are then used to calculate a PEKA score that represents motif enrichment. Alternatively, PEKA offers users an option to calculate enrichment by using all sequence positions without limiting to a subset of relevant positions.

PEKA score conveys the extent of k-mer enrichment at relevant positions around high-count tXn relative to the same relevant positions around low-count, out-of-peak oXn, which represent the intrinsic background of the studied dataset (Methods, Figure 2.1F). PEKA score is a derivative of standard score—it measures the number of standard deviations separating the estimated k-mer occurrence in the foreground (ARtXn) from the mean estimated occurrence around oXn (μ(ARoXn)) (Methods, Figure 2.1F). For further analyses, the motifs are ranked based on PEKA-score in a descending order and the

top n k-mers are selected to visualise their positioning around tXn with occurrence profiles. By default, PEKA separates the top 20 k-mers into up to 5 clusters based on their similarity in sequence and occurrence profiles, and then visualises the profiles of k-mers within each cluster on the same graph (Methods, Figure 2.1G). PEKA also provides a plot that summarises the prevalence and relative positioning of individual clusters, by showing a sum of k-mer occurrences for each cluster (Figure 2.1H).

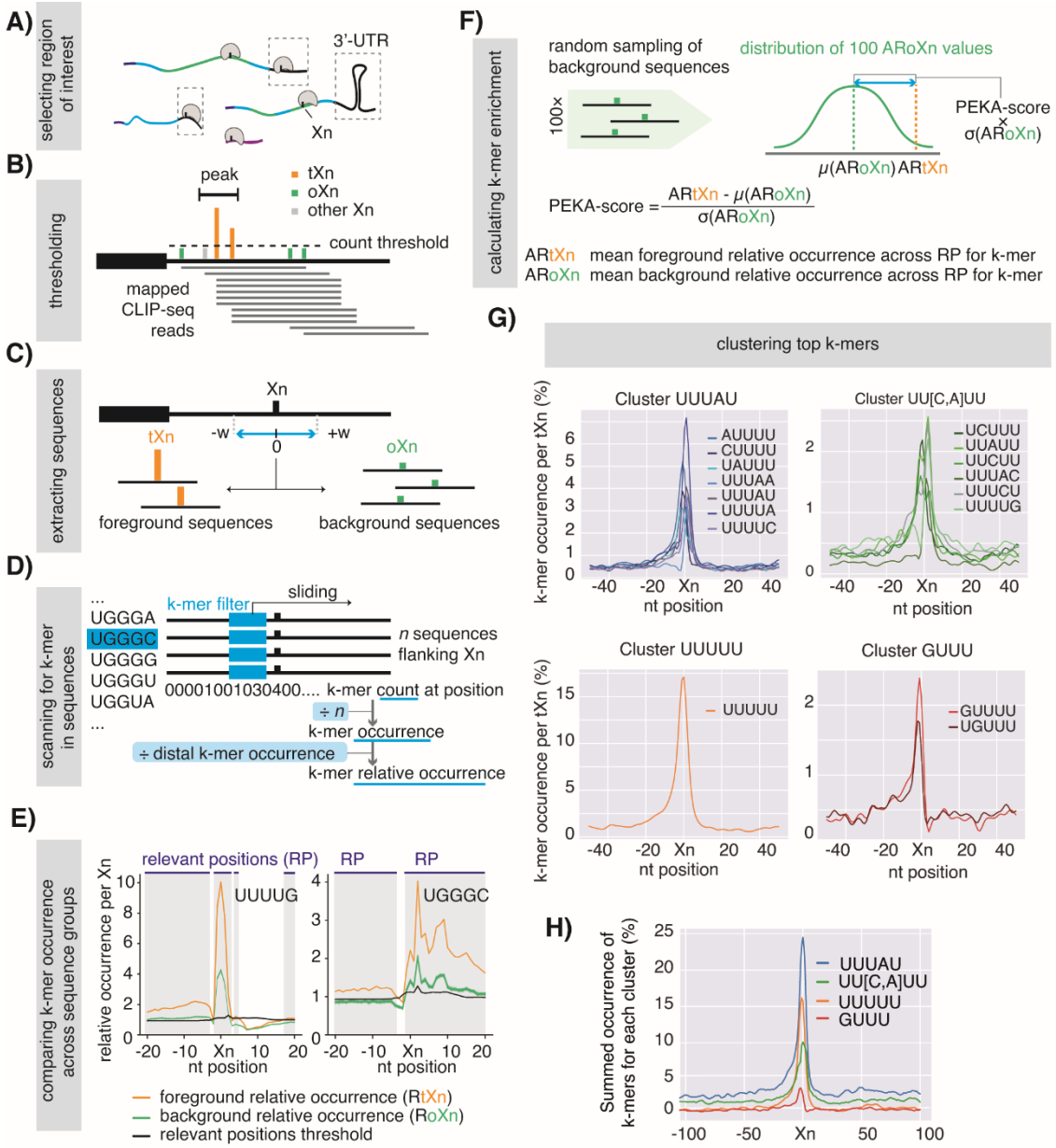


Figure 2.1: Schematic representation of PEKA algorithm.

Continued on next page.

Figure 2.1: *Continued from previous page*

The figure is adapted from our publication (Kuret et al., 2022). (A) PEKA enables motif discovery in distinct transcriptomic regions. (B) PEKA separates crosslink sites (Xn) into tXn (orange) and oXn (green), based on their cDNA score and overlap with user-provided peak regions. (C) Foreground and background sequences flanking tXn and oXn, respectively, are retrieved. (D) Sequences are scanned to record whether a k-mer is present at a particular position. For each group of sequences, k-mer occurrence around Xn is calculated and then converted to relative occurrence (Methods). (E) Positions around Xn where foreground relative occurrence passes the threshold value are considered as relevant positions for enrichment analysis of each k-mer. (F) PEKA score is calculated for each k-mer by comparing k-mer occurrence across the relevant positions around tXn vs randomly sampled oXn (Methods). (G) Top n k-mers with the highest PEKA score are clustered to represent the RBP binding motifs, based on their sequence and occurrence around crosslinks. The panel shows four clusters identified for TIA1 eCLIP in HepG2 cell line. Each cluster is titled with a representative and unique consensus, generated from k-mer alignments. Each line on the plot represents the occurrence of one k-mer around tXn. (H) In addition to individual k-mer clusters, PEKA provides a summary plot, which shows summed k-mer occurrences within each cluster. This plot can be used to compare the quantity and positioning of individual k-mer clusters around tXn.

2.2 PEKA Detects Multiple Binding Modes of RBPs

To study whether PEKA can detect multiple binding modes of RBPs, we investigated enriched k-mers and their distribution around crosslinks, for two RBPs that are known to bind distinct types of motifs—the TAR DNA Binding Protein (TARDBP) and LIN28B. To identify groups of enriched motifs for each protein, we clustered the top 20 k-mers discovered by PEKA in eCLIP data based on their sequence. We visualised the motifs with a heatmap showing their relative occurrences around crosslink sites, which are normalised by k-mer's abundance in a distal region (Methods, Figure 2.2). The visualisation of relative occurrences improves the ability to compare the enriched positions of various k-mers, as regional genomic differences can make less-abundant k-mers difficult to see. This is demonstrated by comparing raw k-mer occurrences and relative k-mer occurrences for LIN28B (Figure 2.2A,D).

The top 20 k-mers identified by PEKA for TARDBP form three groups (GU-motifs, GUAU-motifs, and UGAA-motifs) which are consistent with those previously found to have distinct binding preferences to mutant variants of TARDBP in iCLIP experiments—the YG-containing GU-repeats, the YA-containing GU-repeats, and the AA-containing GU-repeats (Hallegger et al., 2021) (Figure 2.2B). These motifs were also discovered by the mCross method, as shown in Figure 2.3. LIN28 is known to bind the (U)GAU motifs and GAGG motifs (Ustianenko et al., 2018; Wilbert et al., 2012; Yamamoto et al., 2022). PEKA successfully recovered both of these motif groups among the top 20 enriched k-mers (Figure 2.2A). In contrast, mCross recovered only the GAU motifs, indicating its potentially limited sensitivity towards motifs that crosslink poorly (Figure 2.3). PEKA can also detect and visualise complex binding patterns, such as bipartite motifs bound by QKI (Conn et al., 2015; Galarneau & Richard, 2005) (Figure 2.2C), which are not apparent from the PWM-based visualisations provided by mCross (Figure 2.3).

These examples demonstrate that the visualisation of k-mer clusters can offer valuable insights into diverse binding modes of RBPs. To enable the exploration of different binding patterns to the broader research community, we provide heatmaps of the top 40 k-mers

discovered by PEKA for all 223 ENCODE eCLIP datasets on the interactive web interface at https://app.flow.bio/peka/ (Capitanchik et al., 2023; Kuret et al., 2022).

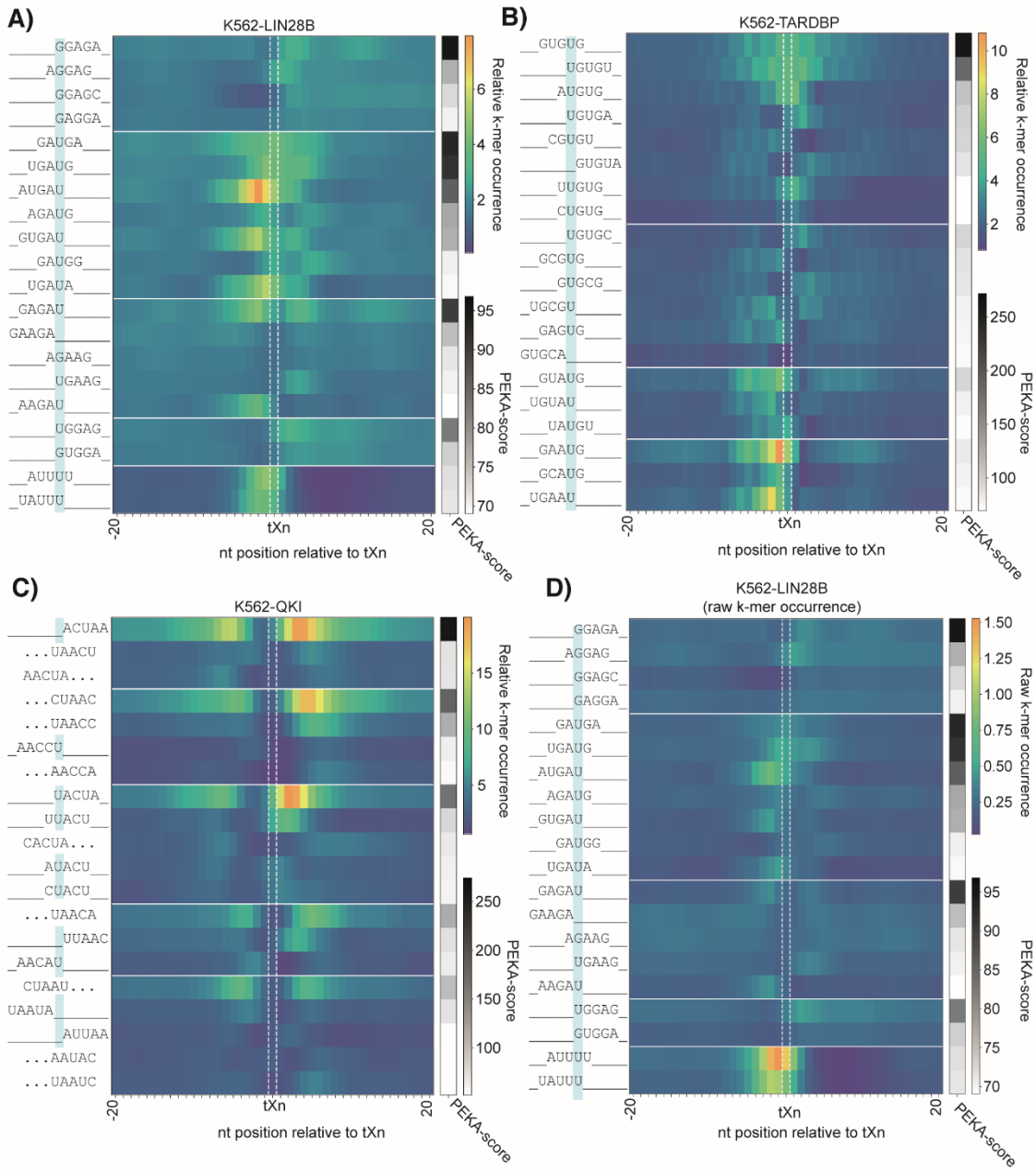


Figure 2.2: PEKA detects different binding modes of RBPs and motifs with complex patterns.

(A-C) Heatmaps of relative k-mer occurrence around tXn for 20 most enriched 5-mers for (A) LIN28B eCLIP, (B) TARDBP eCLIP and (C) QKI eCLIP in K562 cell line. K-mers are clustered based on their sequence and on the left of the heatmaps, their sequences are aligned with the position of relative occurrence maximum. The blue line which spans across labels highlights the most frequently crosslinked nucleotide in the k-mer. (D) Heatmap of raw k-mer occurrence (showing % of crosslinking events that have a motif located at the visualised position) around tXn for 20 most enriched 5-mers for LIN28B eCLIP in K562 cell line. K-mers are clustered in the same manner as described for panels A, B and C.

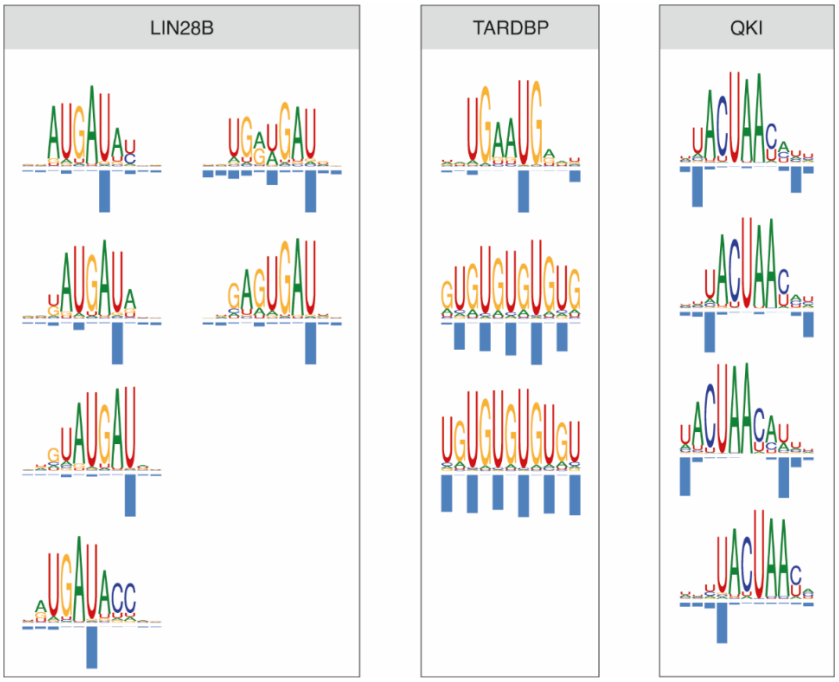


Figure 2.3: Motifs identified by mCross in eCLIPs of LIN28B, TARDBP, and QKI in K562 cell line.

2.3 Benchmarking of PEKA Against Comparable State-of-the-Art Method and *in vitro* Data

A major challenge in motif discovery from CLIP data is the identification of motifs that correspond to the RNA-binding specificity of the studied RBP, as opposed to motifs enriched for other confounding reasons, such as the technical biases of CLIP, associated features of genomic regions and repetitive elements bound by the protein, or motifs bound by other RBPs that may be co-purified if IP stringency was insufficient (Hafner et al., 2021). To evaluate the specificity of enriched motifs, these can be compared with the motifs obtained by *in vitro* methods like RBNS (Lambert et al., 2014) and RNAcompete (Ray et al., 2009). These methods enrich RNAs bound by recombinant RBP in a highly controlled environment, which allows for the examination of its intrinsic RNA-binding affinity. Because *in vitro* methods are not subject to the same biases as CLIP data, they are well-suited for examining the biological specificity of motifs derived from CLIP. We evaluated the motif discovery performance of PEKA by examining the overlap between the highest-ranking k-mers it identified from eCLIP data and the top 20 k-mers that were enriched *in vitro* for the same RBP (Figure 2.4A,B). We also compared the performance of PEKA with that of a previously published mCross method on the same data (Feng et al., 2019). In total, 41 eCLIP datasets were compared for 28 distinct proteins for which both mCross and *in vitro* data were available (Figure 2.4A).

To obtain the data for the analysis, we identified enriched 5-mers in the relevant eCLIP datasets with PEKA; the corresponding mCross enrichment scores were provided by Zhang lab (Feng et al., 2019). For relevant RBPs, we obtained *in vitro* k-mer enrichment scores from published RBNS and RNAcompete datasets. While RBNS and PEKA assessed the enrichment of 5-mers, RNAcompete and mCross evaluated 7-mers. To enable a comparison between methods, the mCross and RNAcompete z-scores were converted from 7-mer to 5-mer scores (Methods).

For each eCLIP dataset, we ranked 5-mers from PEKA and mCross, based on their enrichment. Then, we calculated a "*recall*" metric for each evaluated dataset as a proportion of the top 20 k-mers from *in vitro* data that were recovered among the top n k-mers from the corresponding eCLIP dataset (values of n included 20, 30, 40, 50, 75, 100, and 150). At each n, we plotted the mean and standard deviation of recall across all evaluated eCLIP datasets for PEKA and mCross (Figure 2.4A). This analysis revealed that PEKA performed very similarly to mCross in recovering the top 20 k-mers from *in vitro* data, with slightly higher variability. Both approaches successfully recovered a quarter of the top 20 *in vitro* k-mers within top 20 eCLIP k-mers, indicating high enrichment of RBP-specific biologically relevant motifs. We also evaluated recall (at $n=50$) achieved by PEKA for 16 eCLIP datasets with orthogonal *in vitro* data available, which were not analysed by mCross (Figure 2.4B). For these eCLIPs, PEKA performed well, achieving a recall greater than 0.5 for 11 out of 16 experiments. To further compare the performance of PEKA and mCross, we identified eCLIP datasets, for which the two methods showed a difference in recall (at n=50) greater than 0.2—meaning that one of the methods recovered 20% more of *in vitro* k-mers compared to another. We found that PEKA consistently outperformed mCross for proteins that bind to motifs with high U-content; conversely, mCross performed better for RBPs hat bind motifs with low U-content (Figure 2.4C). It is important to note that *in vitro* methods may fail to capture all the biologically relevant binding modes of RBPs *in vivo*, as the system lacks post-translational modifications of the protein, post-transcriptional modifications of the RNAs, long-range RNA structures and protein co-factors, which might be required for its binding. Nevertheless, they remain useful for evaluating the performance of motif-discovery in CLIP data, as they enable a fair comparison among different tools and provide insight into the relevance of enriched motifs.

This analysis demonstrated that the workflow used by PEKA performed equally well as the workflow used for k-mer enrichment analysis in mCross, despite the differences in the selection of background and foreground sequences, as well as the approach used to calculate enrichment scores. In PEKA, the foreground sequences are extracted around thresholded crosslinks, the background sequences are extracted around low-count out-of-peak crosslinks and the k-mer enrichment is calculated across relevant positions around crosslink sites, increasing sensitivity (Figure 2.1E). Finally, enriched motifs are represented as k-mer clusters (Figure 2.1G). Conversely, mCross uses transcript regions with a high density of crosslinks, i.e. peaks, as foreground sequences to evaluate raw k-mer enrichment, with respect to distal regions—located 450nts up- and downstream of the peak centres. mCross workflow proceeds with normalisation of raw enrichment scores at the level of the entire eCLIP dataset, by subtracting the median z-score of each k-mer across all experiments followed by robust scaling using the median absolute deviation. This ubiquitously reduces the scores of motifs that are commonly enriched across diverse eCLIP datasets—primarily G-rich motifs (Feng et al., 2019). Next, mCross uses normalised scores to assess each dataset for asymmetrically enriched k-mers, which mark the sequence specificity of RBPs. If the dataset lacks asymmetrically enriched k-mers, the motif analysis is terminated. mCross then proceeds to *de novo* motif discovery, using the top 10 asymmetrically enriched k-mers as seeds. Binding motifs are represented as PWMs (Figure 2.3). For this benchmark, we focused solely on k-mer ranking in mCross rather than the final *de novo* motifs, as the latter require cross-dataset normalisation, and reflect only a small portion of the top ranked k-mers, with which the *de novo* motif discovery was initialised.

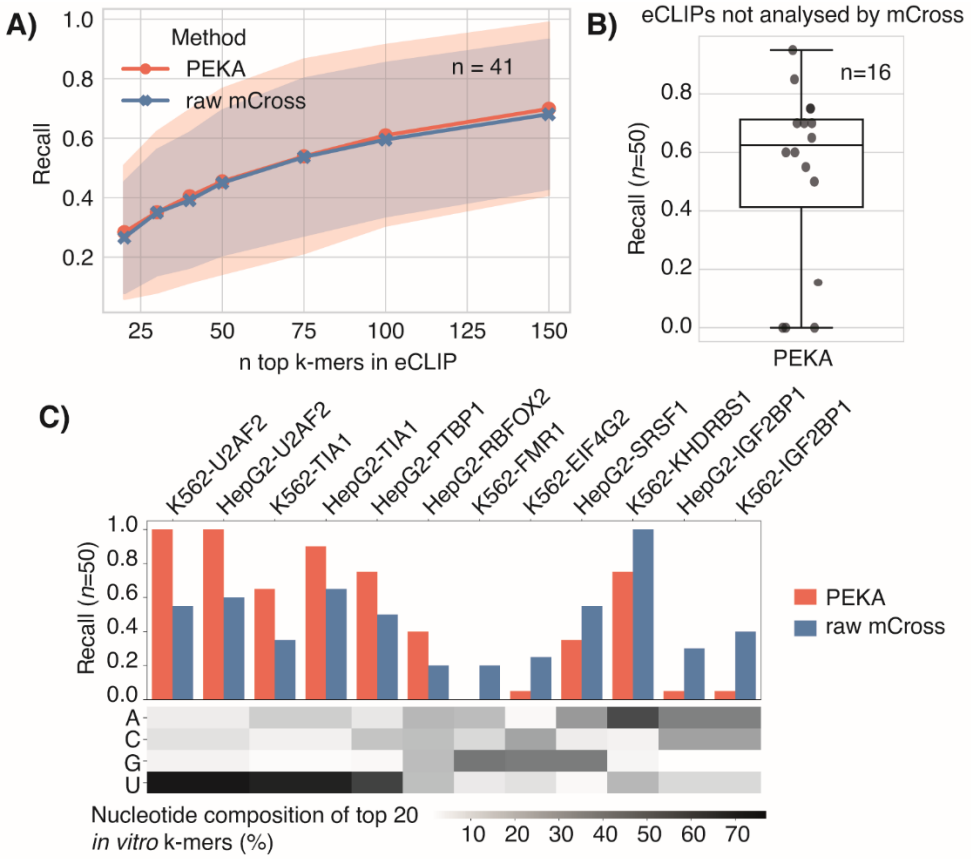


Figure 2.4: Benchmarking motif discovery performance of PEKA on eCLIP data against mCross and *in vitro* data.

(A) Comparison of PEKA and raw mCross in their ability to recover the top 20 k-mers from *in vitro* dataset in corresponding eCLIP data. The lines show the mean recall value at each threshold n and the shaded areas represent the standard deviation. (B) Boxplot of recall values for PEKA analysis of eCLIP datasets that were not analysed by mCross. (C) Bar plot shows the recall of PEKA and mCross for eCLIP datasets in which the absolute difference in recall between the tools was greater or equal to 0.2. Heatmap shows the nucleotide composition of top 20 *in vitro* k-mers for the corresponding eCLIP data.

2.4 Summary

This chapter discussed PEKA; a motif-discovery tool specialised for CLIP data that uses precise positions of crosslink sites to derive motifs. We explained its algorithm design (Figure 2.1), analysed its ability to detect various types of motifs and RBP binding modes (Figure 2.2), and compared its performance with the state-of-the-art method mCross (Figure 2.4).

We compared motifs discovered by PEKA and mCross from eCLIP data to motifs obtained with *in vitro* RNA-binding assays, which are considered the gold standard for understanding intrinsic RBP binding preferences. We found that, on average, PEKA and mCross recovered relevant binding motifs from CLIP data at a similar rate (Figure 2.4A). When we analysed the performance of both algorithms at the level of individual RBPs, we discovered that PEKA performed better for RBPs that bind to U-rich motifs, while mCross performed better for RBPs that bind motifs lacking Us (Figure 2.4). These findings suggest

that while the approach of using low-count crosslinks for background modelling employed by PEKA leads to the discovery of relevant motifs, it does not outperform the mCross method across the board. Instead, PEKA and mCross each have their strengths depending on the RBP's sequence specificity.

Despite the lack of performance improvement over mCross, PEKA still presents several notable enhancements. For instance, PEKA's representation of enriched motifs as k-mer clusters allows the visualisation of more complex enrichment patterns, as exemplified by QKI (Figure 2.2C), and the detection of various binding modes of RBPs, as exemplified by LIN28B (Figure 2.2A). Additionally, PEKA enables motif analysis in various genomic regions (Figure 2.1A) and its source code is freely accessible from GitHub and Bioconda.

Chapter 3

Specificity of Protein-RNA Interactions Observed by CLIP

In this chapter, we apply PEKA to a meta-analysis of CLIP data, comparing enriched motifs across different RBPs and variants of the CLIP method. This comprehensive study, considering *in vitro* motifs, RBP structure, genomic regions, and markers of CLIP data, produced biological insights into the sequence specificity of RBPs. We establish that RBPs without canonical binding domains and with more IDRs and compositional biases have lower sequence specificity towards linear sequence motifs and a higher enrichment of common motifs.

Moreover, we gain technical insights into how methodological biases of CLIP and characteristic sequence preferences of genomic regions impact motif discovery. We demonstrate that motifs enriched by different CLIP methods are susceptible to certain sequence preferences, which relate to the method's technical biases and sequence properties of genomic region that contains the majority of thresholded crosslink sites. We observe that the enrichment of common motifs is related to indicators of data sensitivity, such as the RBP's ability to crosslink and the number of thresholded crosslinks. Finally, the text and figures in the following sections are adapted from our publication (Kuret et al., 2022).

3.1 Insights into the Specificity of Motifs Detected by Different Variants of CLIP Method

PEKA enables motif analysis in any type of CLIP data, including both putative crosslink sites resulting from RBP binding and non-specific sites that represent technical noise. In this section, we compare enriched motifs identified by PEKA for eCLIP, iCLIP, and PAR-CLIP to understand how they differ across different CLIP methods and whether datasets produced by each variant exhibit common motif enrichment patterns. We performed the analysis in two parts: first, we conducted a case study of TIA1, an RBP with well-studied motif binding preferences for which all three CLIP variants were available; secondly, we compared enriched motifs identified by PEKA across multiple RBPs for eCLIP, iCLIP, and PAR-CLIP datasets to understand how they differ across different CLIP methods and whether datasets produced by each variant exhibit common motif enrichment patterns. We analysed the enriched motifs in the context of the corresponding *in vitro* experiments, the content of thresholded crosslinks in specific genomic regions, and the content of crosslinks in repetitive elements to understand how these properties of CLIP data relate to motif enrichment.

3.1.1 A case study of TIA1

To get preliminary insights into the motifs that PEKA discovers in different variants of the CLIP method, we analysed the sequence and occurrence distribution of top 40 5-mers enriched for the well-studied protein TIA1 in eCLIP, iCLIP and PAR-CLIP (Figure 3.1). TIA1 is known to bind U-rich sequence motifs (López de Silanes et al., 2005), which are indeed enriched among the top 20 k-mers in *in vitro* RBNS experiment (Figure 3.1A). We evaluated the recovery of these top 20 *in vitro* motifs in different CLIP datasets, by calculating the recall metric; this showed that for TIA1, PEKA recovered relevant motifs in all three CLIP variants; however the recall was slightly lower in PAR-CLIP, compared to eCLIP and iCLIP (Figure 3.1A).

The 5-mer profiles for TIA1 eCLIP, iCLIP, and PAR-CLIP show the most enriched motifs to be U-rich (Figure 3.1B,C,D), in agreement with the known sequence specificity of TIA proteins. U-rich motifs are most enriched directly at crosslink sites, consistent with the sequence preferences of UV-crosslinking, which occurs predominantly on uridines (Hafner et al., 2021; Knörlein et al., 2022). In eCLIP of TIA1 G-rich motifs located primarily downstream of the crosslink sites are also detected (Figure 3.1B). We found that G-rich motifs were also detected by the orthogonal mCross method for the same dataset (TIA1 in HepG2 cells). Notably, among the top 20 k-mers ranked by the raw mCross score, 11 were U-rich and 9 were G-rich (Kuret et al., 2022); see Additional file 6 of the cited study. This confirms that G-rich motifs are indeed enriched in regions with high density of crosslink sites in eCLIP of TIA1. In iCLIP of TIA1 we observe the CGGA motifs, the GUA motifs and the A-rich motifs enriched across broader regions around crosslink sites (Figure 3.1C). The heatmap visualisations in Figure 3.1 clearly demonstrate that in each CLIP experiment, k-mers with similar sequences tend to have similar positional profiles around crosslink sites and that PEKA can report these distinct k-mer groups. However, it is unclear whether the non-U-rich motifs identified in iCLIP and eCLIP are biologically relevant, i.e., linked to the binding of TIA1 to the RNA in tested conditions.

Non-overlapping positional patterns of different motif groups can represent different modes of binding by the RBP of interest or motifs of RBP co-factors that bind RNA in tandem with the target RBP—leading to association of different motif types; however, they can also result from a non-RBP-specific source, such as the co-purification of other RBPs, due to insufficient stringency of IP. To evaluate the relevance of newly discovered motifs, they can be compared to those recovered by other CLIP methods under the same conditions; if multiple CLIP methods manage to detect the same motifs, they are likely biologically relevant. The non-U-rich groups identified for TIA1 do not overlap between different CLIP methods, indicating that they might be representative either of cell-type-specific binding patterns of TIA1—as the experiments were performed in different cell types; or they could reflect non-specific artefacts. To better understand the biological relevance of these motif groups, they can be compared to other experiments targeting different RBPs but performed with the same CLIP method. If the identified motifs are commonly enriched in many datasets of distinct RBPs, their detection is likely not RBP-specific.

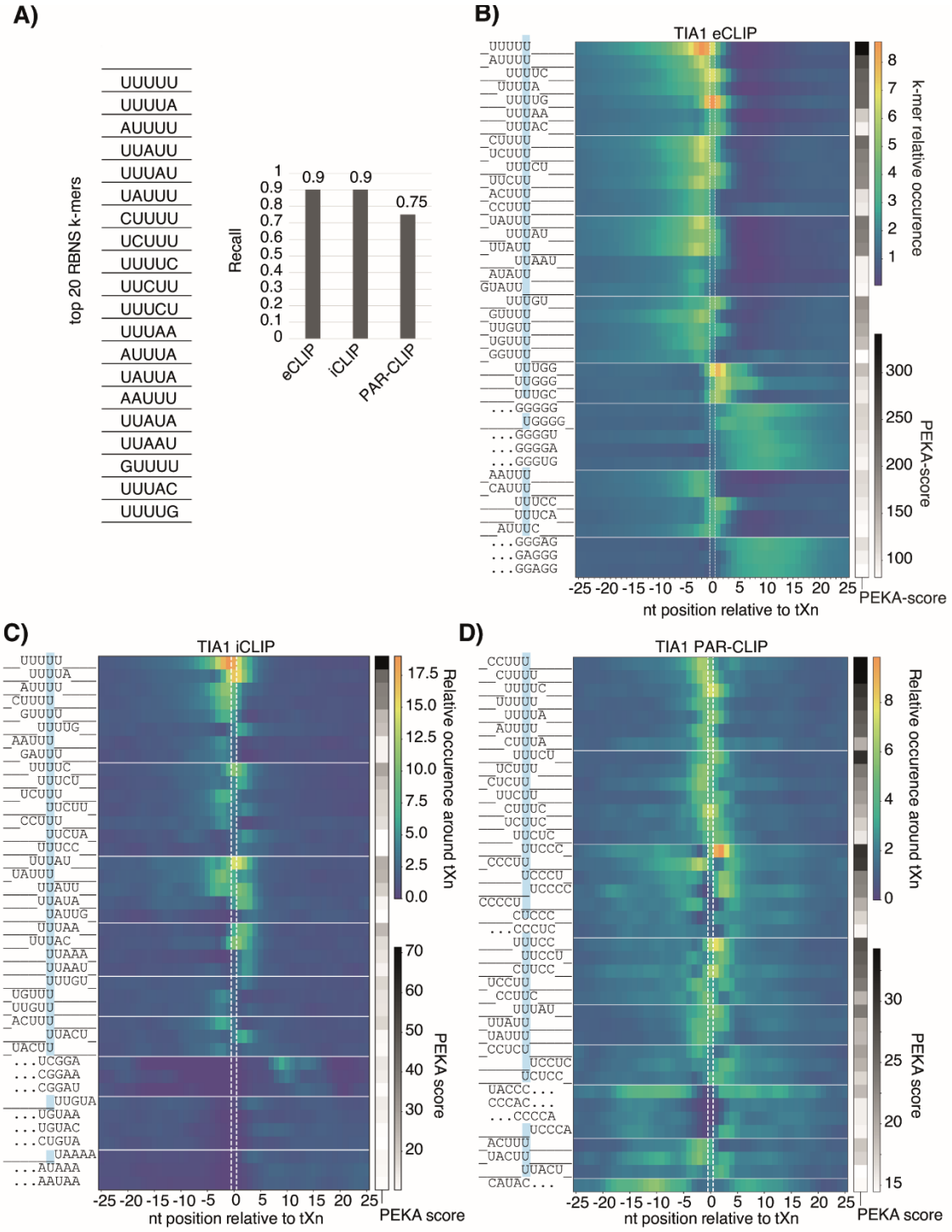


Figure 3.1: Binding specificity of TIA1, as detected in *in vitro* experiments and different CLIP methods.

(A) Top 20 enriched 5-mers for TIA1 in RBNS (left) and recall of TIA1 CLIP experiments, performed with different variants of CLIP methods (right). The recalls correspond to CLIP experiments shown in panels B-D. (B-D) Heatmaps of relative k-mer occurrences around tXn for 40 most enriched 5-mers for B) TIA1 eCLIP (HepG2 cells), C) TIA1 iCLIP (HeLa cells), D) TIA1 PAR-CLIP (HEK293 cells).

3.1.2 A meta-analysis of enriched motifs in diverse eCLIP, iCLIP, and PAR-CLIP datasets

To learn more about the general motif preferences of different CLIP protocols, we performed a more detailed comparison of PEKA with the results obtained by *in vitro* method RBNS, which are available along with eCLIP data for 21 RBPs, iCLIP data for 4 of these RBPs, and PAR-CLIP data for 9 RBPs. PEKA analysis was performed in the "protein-coding gene" region, which combines intron, CDS, and the UTRs. Repeat sequences were filtered out in motif detection. Visualisation of k-mer ranking across groups, generated by sequence-based clustering, forms a unique binding signature for each dataset (Figure 3.2), revealing informative variations in the ranking of top k-mers between data produced by these three CLIP methods.

In the case of TARDBP, FUBP3, KHSRP, TIA1, HNRNPC, HNRNPL, PCBP1, and PCBP2, eCLIP as well as most of the iCLIP and PAR-CLIP experiments show high agreement with RBNS (i.e., recall, Figure 3.2). The G-rich motifs, which were previously observed to be enriched for TIA1 eCLIP, but not iCLIP or PAR-CLIP (Figure 3.1), are also enriched in RBFOX2 and IGF2BP1/2 eCLIP, but not in RBNS, or IGF2BP1/2 PAR-CLIP. In the case of IGF2BP1/2, additional divergence can also be attributed to the enrichment of C-rich motifs in eCLIP, which RBNS and PAR-CLIP do not exhibit. In addition to IGF2BP1/2, eCLIP experiments for PUM1, SFPQ, RBM22, and EIF4G2 also showed poor agreement with RBNS data. In the case of PUM1 and IGF2BP1/2, the data from PAR-CLIP are in much better agreement with RBNS than eCLIP. Instead of the expected motifs, G-rich motifs were enriched in the PUM1 eCLIP, suggesting that these motifs tend to be enriched when an eCLIP experiment fails to identify the expected signal (Figure 3.2). Considering the known similarity of motif specificity of PUM1 and PUM2 (Spassov & Jurecic, 2003), we compared the PUM2 eCLIP experiment with the *in vitro* data of PUM1, which showed high agreement, as reported previously (Van Nostrand et al., 2020). This analysis demonstrates a substantial variance in the reliability of eCLIP datasets, and the value of CLIP meta-analyses to identify the datasets that are likely to be the most reliable for further studies.

We found that when CLIP datasets differ in enriched motifs, they also tend to differ in the regional distribution of crosslink sites (Figure 3.2). Specifically, compared to eCLIP and iCLIP, PAR-CLIP generally exhibits a higher proportion of tXn in the 3′-UTR relative to introns, and a lower coverage of repetitive elements. The increased proportion of tXn in 3′-UTRs, observed in PAR-CLIP, is likely a result of the experimental design, as PAR-CLIP experiments typically target exogenous FLAG-tagged proteins, rather than endogenous proteins (Mukherjee et al., 2019). While this approach ensures comparable protein expression between experiments and allows for more stringent conditions of protein purification, the supraphysiological expression of exogenous proteins often leads to their accumulation in the cytoplasm, which could account for the overall higher proportion of tXn in the 3′-UTRs, compared to intronic regions.

In addition to observing variable motif patterns and crosslink distribution between different CLIP methods, we also found variations in the regional distribution of tXn between eCLIPs of homologous proteins PUM1 and PUM2. Specifically, we found PUM1 to have a lower proportion of tXn in the 3′-UTR and a higher proportion of tXn in CDS and 5′-UTR. Thus, a combined analysis of several CLIP features, such as motif enrichment and regional binding, may be particularly valuable for data quality assessment, and for understanding the potential generic biases of each CLIP variant.

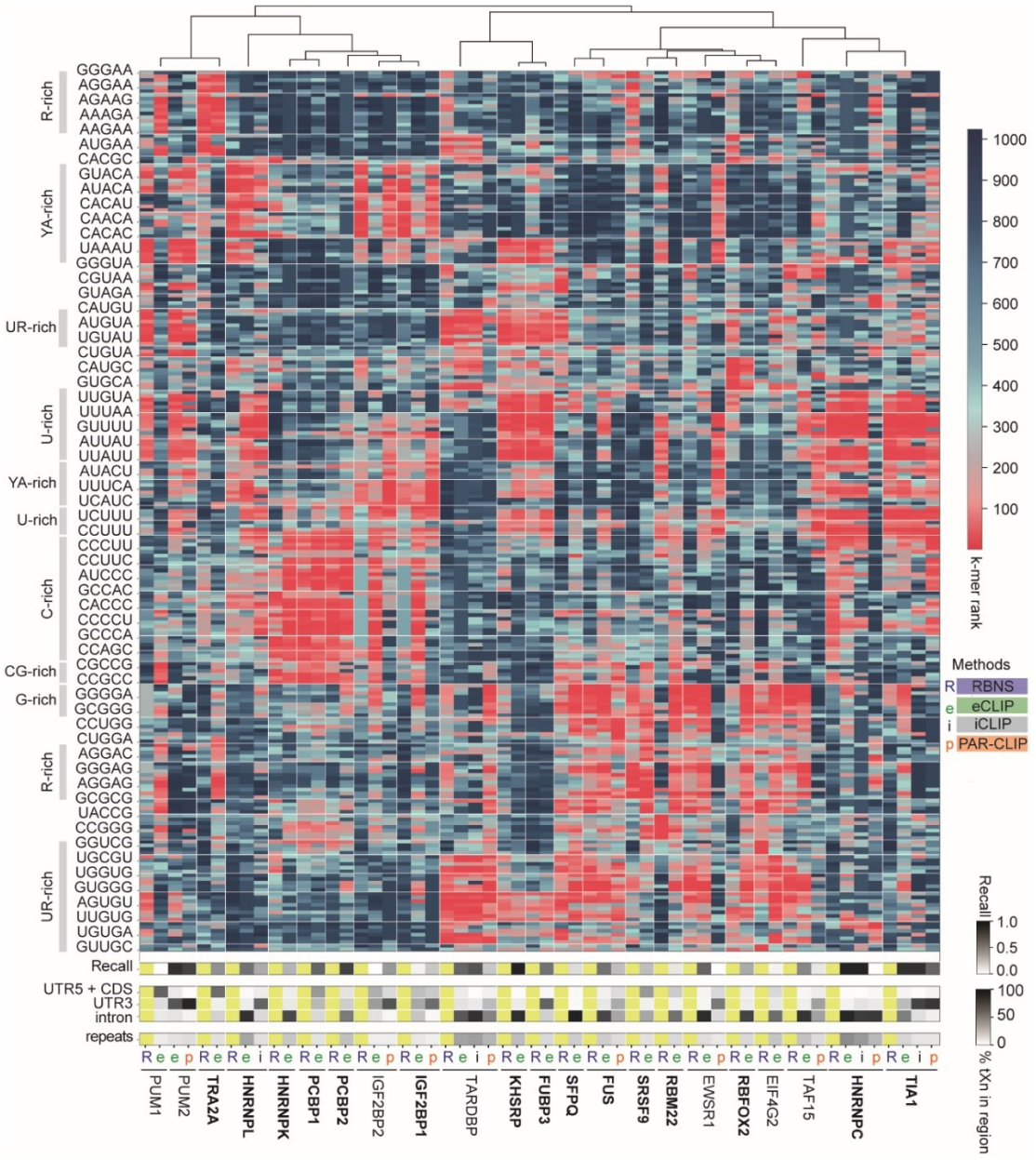


Figure 3.2: Comparison of CLIP methods against RBNS.

Heatmap shows the rank order of k-mer PEKA scores for each CLIP dataset, and of RBNS z-scores. The scale on the top right shows these values spanning from 1 to 1024. K-mers selected for this heatmap (n=230) ranked among the top 10 in eCLIP, PAR-CLIP, iCLIP, or RBNS for any of the RBPs shown. K-mers were clustered based on their sequence and ordered using hierarchical clustering of median dataset ranks within clusters. On the left, prominent k-mer sequence features are highlighted across clusters. Below, the recall heatmap shows how much the top motifs in the dataset correspond with the RBNS data. The third heatmap shows the percentage of tXn derived from different transcript regions, and the fourth heatmap shows the percentage of tXn found in repeat sequences. eCLIP experiments in bold are in HepG2, others in K562 cell line. In case two eCLIP datasets for the same RBP were available, the one with higher recall is shown.

3.2 Insights into the Technical Biases of CLIP

The following sections describe the analysis of sequence biases associated with crosslink sites in eCLIP and PAR-CLIP experiments, and how they are minimised by motif-discovery tools PEKA and mCross.

3.2.1 eCLIP

To illustrate how PEKA controls for technical biases in CLIP, we compared the k-mer rankings between PEKA, mCross, and a local approach that accentuates sequence biases associated with crosslink sites. The local approach examines motif occurrence in a narrow window around tXn sites (−3...3nt), normalised by the average occurrence within distal windows (−150…−100 and 100…150nt around tXn); PEKA, conversely, uses k-mer occurrence at relevant positions around tXn as foreground and k-mer occurrence at relevant positions around oXn as background to calculate motif enrichment. Therefore, the local approach serves to emphasise motifs enriched at crosslink sites, while trying to reduce the contribution of non-specific regional genomic sequences by normalising with k-mer occurrences in a distal window. PEKA, however, does not indiscriminately emphasise motifs focused directly on the crosslinks site, but uses positions around crosslinks, where the enrichment of a specified k-mer is high; it then normalises k-mer occurrence at these positions with the occurrence around low-count crosslink sites, simultaneously reducing the sequence biases of genomic regions and crosslink sites. We compared the performance of the local approach and PEKA, by analysing recall across 57 eCLIP datasets (Figure 3.3A). We found that the enrichment calculation used in PEKA performed significantly better than the local approach. This indicates that the motifs present in a narrow window around crosslink sites are not necessarily relevant to RBP binding, and that relevant motifs are hard to recover at these positions by using distal window for normalisation. In contrast, the selection of relevant positions and normalisation with the intrinsic background used by PEKA recovers more RBP-specific motifs—as demonstrated by higher agreement with *in vitro* data.

To understand why certain motifs are more or less likely to be detected in eCLIP data, compared to *in vitro* data, we compared k-mer enrichments obtained by the local approach, PEKA, mCross or *in vitro* across a subset of 41 eCLIP datasets (representing 28 distinct RBPs) that had both *in vitro* and mCross data available. We identified groups of k-mers that were differentially ranked in each method relative to *in vitro* data (confidence interval $> 95\%$ and a fold-change greater than 1.5 or less than 0.66) (Figure 3.3A, Methods). Because the local approach was designed to accentuate the biases associated with crosslink sites, it produced the highest number of k-mers that were differentially enriched in eCLIP vs *in vitro* data (~27%); PEKA, conversely, produced the lowest number of such differentially enriched k-mers (~10%) (Figure 3.3B,D,E).

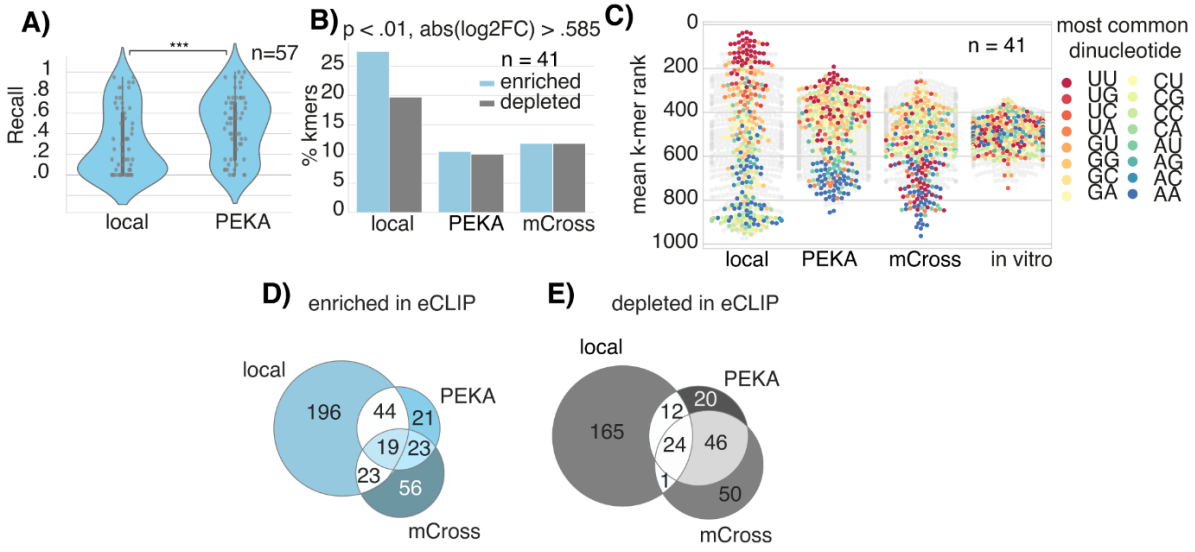


Figure 3.3: Differential enrichment of motif groups in eCLIP compared to *in vitro* data.

(A) The recall of local approach and PEKA for 57 eCLIP datasets ($p < 0.001$, paired t-test with Bonferroni correction, alpha = 0.05). (B) Percentages of the disproportionately enriched and depleted k-mers in eCLIP, compared to *in vitro,* for local, PEKA, and mCross approaches. (C) Mean k-mer ranks across RBPs in each approach. K-mers that contain two or more of the same dinucleotide are coloured by their most common dinucleotide and other k-mers are shown in grey. K-mer ranks in the mCross approach were assigned based on raw scores. (D,E) Venn diagrams show the overlap across groups of 5-mers with significant difference in ranking between *in vitro* and eCLIP, analysed with PEKA, raw mCross, and local approach (enriched k-mers are on the left and depleted k-mers are on the right).

To investigate the general sequence characteristics of differential k-mers, we visualised their mean rankings across all evaluated datasets for each approach. We coloured the k-mers according to their most common dinucleotide, to indicate the sequence preferences of differential k-mers (Figure 3.3C). This revealed that the local approach was predominantly enriched for U-rich k-mers, consistent with sequence biases of UV-crosslinking (Hafner et al., 2021; Knörlein et al., 2022). While the ranking of these k-mers was strongly decreased by PEKA, they remained among the highest-ranking motifs. In mCross, however, the U-rich k-mers were surprisingly among the lowest ranking k-mers. In Figure 2.4, we showed that PEKA consistently outperformed mCross for RBPs that bind to U-rich motifs. This difference in performance could stem from the overall depletion of these motifs in mCross, which likely occurs because U-rich motifs are commonly enriched in a narrow region around crosslink sites, as exemplified by the local approach and analysis of binding motifs of TIA1 (Figure 3.3C, Figure 3.1); the analysis of k-mer enrichment across broad peak regions, performed by mCross may favour repetitive and more diffused motifs, rather than short motifs, centered narrowly on crosslinks.

Besides disproportionately enriched k-mers, we also observed disproportionately depleted k-mers with clear sequence biases. The local approach was strongly depleted for CU, CG, and CC dinucleotides, which was efficiently corrected by PEKA and mCross where such k-mers rank at levels comparable to *in vitro* (Figure 3.3C). Interestingly, AA-rich motifs are depleted in all CLIP analysis approaches as compared to the *in vitro* data, with the strongest depletion seen in mCross (Figure 3.3C). These k-mer sequence

imbalances in eCLIP data demonstrate the importance of assessing global trends of enriched motifs across large numbers of RBPs to understand both the biases of experimental and computational approaches, as well as the true differences in the binding preferences of RBPs between *in vitro* and *in vivo* conditions.

19 k-mers were differentially enriched in eCLIP by all three analytic approaches (the UUCG group, Figure 3.4A) and 24 were differentially depleted (CAUA group, Figure 3.4B). Interestingly, these motifs are evenly enriched up to 150nt around tXn sites as compared to oXn sites (Figure 3.4A). One of the proteins with the highest ranking of these k-mers (in eCLIP) is the DEAD-box helicase DDX3X, which is a major regulator of cellular RNA condensates and itself contains IDRs with strong condensation propensity (Saito et al., 2019). A study reported that DDX3X binds in the vicinity of a motif composed in large part of CG and CGU subsequences, similar to the k-mers found in the UCG-group (Calviello et al., 2021). It could be speculated that such patterns are better detected by CLIP than *in vitro* binding data due to their need to assemble binding-region condensates on long RNA regions as was observed for TDP-43 (Hallegger et al., 2021).

Both PEKA and mCross use a strategy to control for the technical biases of CLIP, and it is thus reassuring that their differential motifs are more similar to each other than to the local approach (Figure 3.4A,B, Figure 3.3D,E). To further understand the unique features of each approach, we examined the motifs that were enriched or depleted uniquely in each approach, or in both approaches (Figure 3.3D,E). This showed that differential motifs identified either by mCross and PEKA tend to be broadly enriched or depleted at over 100-nt region around crosslink sites, whereas as expected, those identified in the local approach are enriched or depleted directly at crosslink sites (Figure 3.4A,B). We also compared k-mer rankings of differential motif groups to their ranking in *in vitro* data, which revealed the depletion of CAUA group to be most prominent in all CLIP analysis approaches and the depletion of the CCCC group to be most prominent in the local approach (Figure 3.4C,D).

Importantly, for all motif groups, we show that PEKA can identify such motifs as top-ranking for the RBPs that show strong enrichment (Figure 3.4A,B). For example, even for the CAUA group motifs that are found depleted in eCLIP by all approaches, compared to *in vitro* data, KHDRBS1 is correctly identified as a strong binder of the motif, with a broad enrichment pattern seen around thresholded crosslink sites (Figure 3.4B). Moreover, PEKA can identify enriched motifs for hnRNPL, TRA2A, and PABPN1 even though they are depleted directly at crosslink sites (Figure 3.4B). Thus, despite generic imbalances observed in the motifs enriched across all CLIP datasets and a great variety of their positional enrichment patterns, PEKA can identify all types of motifs when they mediate high-affinity RBP binding.

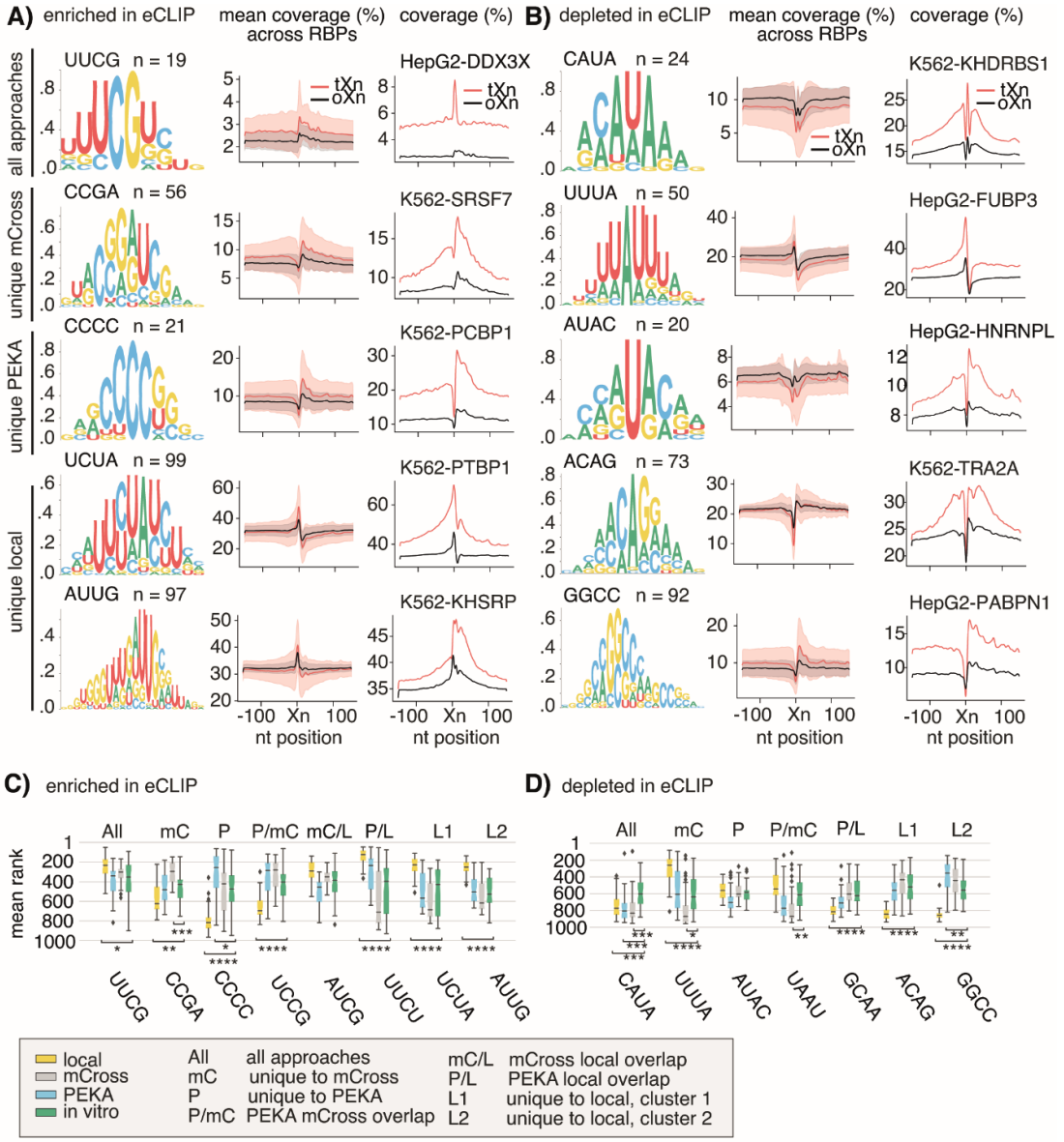


Figure 3.4: Characterisation of motif groups that are differentially enriched in eCLIP compared to *in vitro*.

Continued on next page.

Figure 3.4: *Continued from previous page.*

(A,B) An overview of the differentially enriched (A) and depleted (B) 5-mers in eCLIP analysis approaches relative to *in vitro* data. In each panel, k-mer groups are arranged from top to bottom as follows: k-mers that are differentially ranked in all approaches, only in raw mCross, only in PEKA, or only in local approach. For local approach k-mers were split into 2 clusters based on their sequence, due to high number and diversity of k-mers in this group. For each group we show its k-mer logo (Methods), and two line plots: the left line plot shows the distribution of mean motif group coverage around tXn (red) or oXn (black) within the protein-coding gene region across the 57 eCLIP datasets with available *in vitro* data (shaded areas show SD); the right line plot shows the distribution of motif group coverage around tXn (red) or oXn (black) within the protein-coding gene region for an RBP that ranked highly in both eCLIP and *in vitro* data for the corresponding motif group. (C,D) Boxplots show mean k-mer ranks across evaluated RBPs in *in vitro* data and in approaches used to analyse eCLIP data for the enriched (C) and depleted (D) motif groups. The legend is shown in a bracket below. For each motif group, we compared the distribution of mean ranks in CLIP analysis approaches with *in vitro* data to assess the significance of difference in the mean ranks with Welch's *t*-test. We marked only the comparisons with $p < 0.05$.

3.2.2 PAR-CLIP

To investigate the biases that are associated with crosslink sites in transition-based CLIP variants, we expanded our analysis to 19 PAR-CLIP datasets for which *in vitro* data was available. In PAR-CLIP, crosslinking is performed at a higher wavelength and with the addition of a photoreactive ribonucleoside analogue 4-SU, which, upon sequencing, allows for the precise locations of crosslink sites to be identified from T-to-C transitions. As for eCLIP data, we applied PEKA and a local approach to PAR-CLIP experiments and compared the recall of both approaches. In contrast to eCLIP, we observed only a slight improvement of recall in PEKA compared to the local approach; the change was not significant (Figure 3.5A). Furthermore, the recall of PEKA and local motifs obtained from PAR-CLIP data was generally lower compared to the RBP-matched eCLIP datasets (Figure 3.5B), however, the difference was not statistically significant, likely due to high variance in recall and small sample size (n=12).

We then analysed the proportion of k-mers that exhibit significant differences in ranking between PAR-CLIP and *in vitro* data and again found that PEKA greatly reduces the number of differential k-mers as compared to the local approach (Figure 3.5C). When we analysed k-mer ranking with respect to their sequence composition, we observed relatively similar trends as in eCLIP, with U-rich k-mers being highly enriched and C-rich k-mers depleted in the local approach, while the effect is diminished in PEKA k-mers. Interestingly, the bias against AA k-mers that was observed in eCLIP data is not apparent in PAR-CLIP, and instead the bias against GG-containing k-mers is seen in PAR-CLIP (Figure 3.5D).

Together these results suggest that while PEKA does reduce crosslink-associated biases of PAR-CLIP experiments, the method performs worse for transition-based CLIP methods, compared to truncation-based CLIP methods. This drop in performance could be explained by lower sensitivity of PAR-CLIP, compared to truncation-based methods; PAR-CLIP experiments identify fewer crosslink sites, but the level of background is also lower. This could lead to the assignment of crosslinks representative of target RBP binding to background and incur a depletion of relevant motifs upon enrichment computation; nevertheless, the local approach without oXn normalisation does not improve recall, compared to PEKA, and has more prominent sequence biases, therefore it is unclear whether the normalisation with oXn is detrimental to performance.

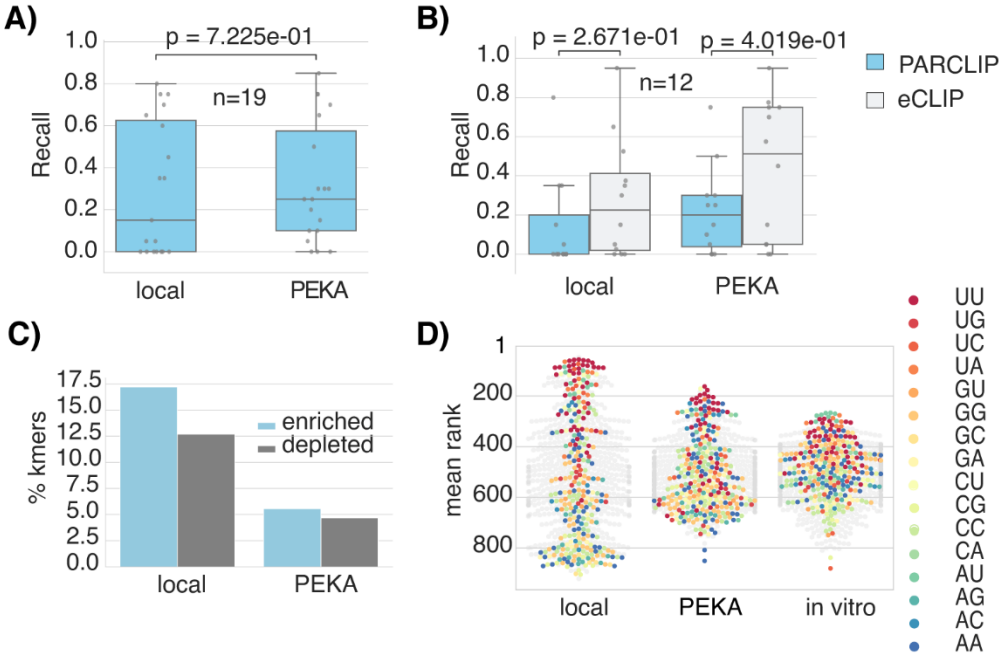


Figure 3.5: Differential enrichment of motif groups in PAR-CLIP compared to *in vitro* data.

(A) Recall achieved by PEKA or local approach across 19 PAR-CLIP datasets for which *in vitro* data was available (p-value is indicated on the plot; paired t-test with Bonferroni correction, alpha = 0.05). (B) Recall for 12 RBPs that had available eCLIP and PAR-CLIP data, obtained by the local approach or PEKA. For RBPs where eCLIPs were available for both cell lines, the mean recall across the experiments is shown. Mann-Whitney-Wilcoxon test was used to calculate p-values, indicated on the plots. (C) Percentage of k-mers with significantly differential ranking (Welch's t-test $p < 0.01$ and a fold-change greater than 1.5 or less than 0.66) between PAR-CLIP and *in vitro* for PEKA and local approach. Differential ranking was assessed on 19 PAR-CLIP datasets for which *in vitro* data was available. (D) Mean k-mer ranks across RBPs (n = 19) in PEKA and local approach. K-mers which contain two or more of the same dinucleotides in their sequence are coloured by their most common dinucleotide and other k-mers are shown in grey.

3.3 Peak Filtering by External Background Yields Limited Benefit in Motif Discovery from eCLIP Data

PEKA was developed for general use on any type of nucleotide-resolution CLIP data, and therefore, it models motif enrichment relative to the intrinsic background of a single dataset, without the need for additional data to model extrinsic background. Nevertheless, we wished to understand if the performance of PEKA on eCLIP data would improve if the input peaks were determined considering the extrinsic background signal. The use of non-RBP-specific controls for identification of RBP-specific peaks is common practice in eCLIP data analysis; eCLIP does not visualise IPed RNP complexes on the gel, prior to their isolation from the membrane, and instead controls for potential co-purification of other RBPs by generating the size-matched input (SMInput) control. SMInput is produced by separating the lysate on the gel, without IP, and isolating the RNPs from the region, that matches the size of the region for RNP isolation in eCLIP samples. The ENCODE consortium provides narrowPeaks, which are generated by retaining only those peaks with a significant enrichment of reads over the SMInput control, which is expected to decrease the extent of extrinsic background in the data (Van Nostrand et al., 2020). We ran PEKA using only the crosslink sites located within the narrowPeaks as the foreground and compared the enriched motifs to those obtained by our Clippy peak-calling approach that does not use extrinsic background.

We observed no overall improvement in motif specificity as compared to our Clippy peak calling approach that does not include SMInput analysis (Figure 3.6A). A major increase in agreement with *in vitro* data when using narrowPeaks was found only for a couple of RBPs, the most prominent being IGF2BP1, but decreased agreement was seen for other RBPs, with the most prominent effect observed for hnRNPC. We find that the approach used to derive narrowPeaks generally results in an increased adenosine content of enriched motifs (Figure 3.6B), which can in turn lead to decreased content of other nucleotides, such as uridine and cytidine. As a result, the use of narrowPeaks tends to improve motif specificity for RBPs that show binding to A-rich motifs by *in vitro* data but decrease specificity for proteins that bind to motifs that do not contain adenosine (Figure 3.6A). To better understand this phenomenon, we visualised motif coverage of the top 20 *in vitro* k-mers for hnRNPC (Figure 3.6C) in hnRNPC eCLIP and SMInput in both cell lines. We found ~25% of cDNA start sites in SMInput have these k-mers enriched around them, in comparison to ~50% (HepG2 cells) or ~35% (K562 cells) in eCLIP (Figure 3.6D), with a very similar profile around crosslink sites in both cases, indicating that hnRNPC likely contributes to a major part of the foreground signal in its SMInput control. Despite eCLIP having a higher proportion of these motifs enriched at crosslink sites, using SMInput as a control for hnRNPC eCLIP likely results in the foreground signal becoming erroneously assigned to the background, precluding the identification of relevant binding sites, an issue that likely affects many other datasets. This could explain the lack of general improvement in motif specificity when using narrowPeaks.

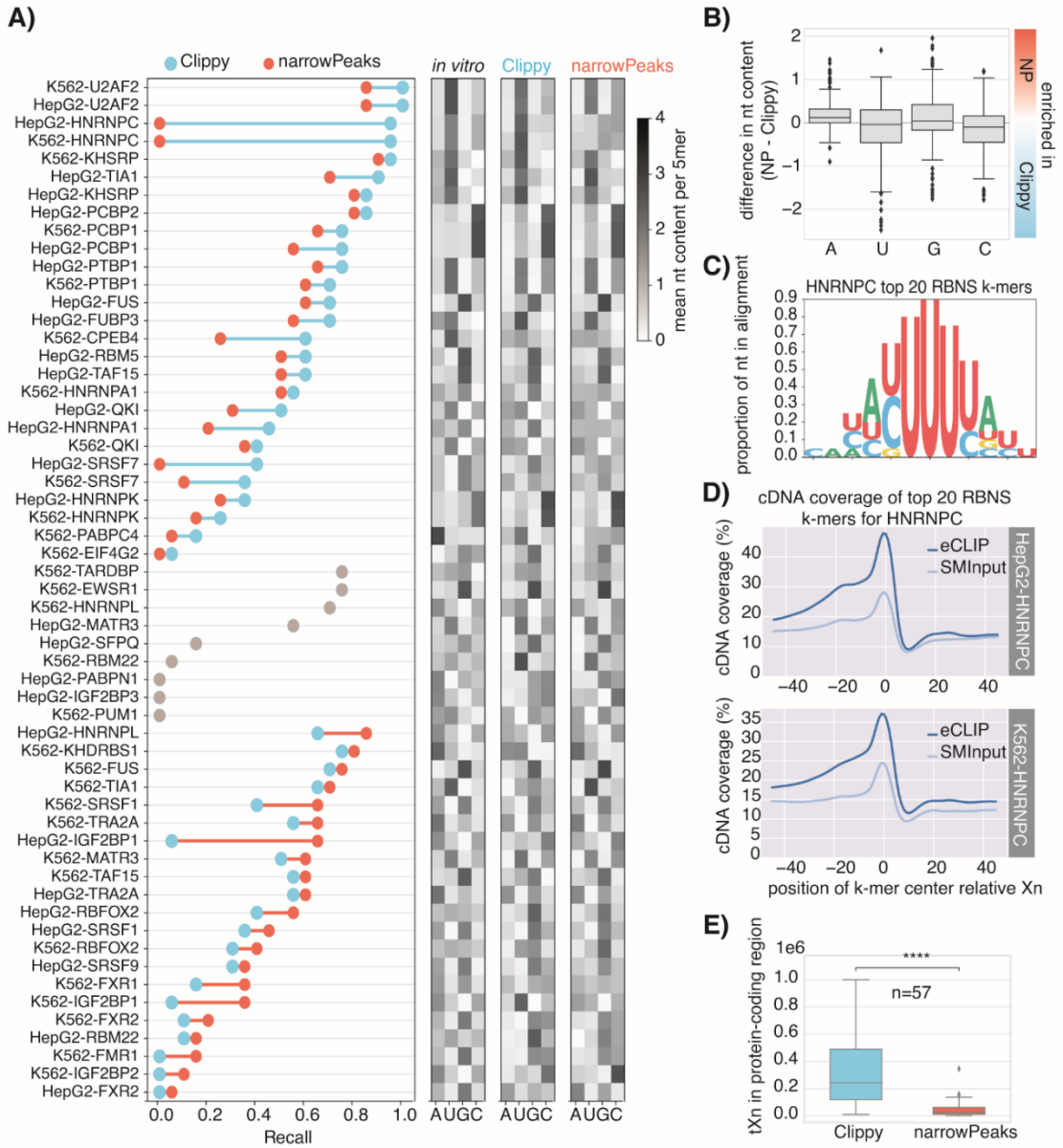


Figure 3.6: Influence of size-matched input controls on motif discovery.

(A) The graph shows recall for 57 eCLIP datasets for which orthogonal RBNS or RNAcompete data was available. The datasets were processed with PEKA, using either Clippy peaks or eCLIP narrowPeaks from merged replicates, downloaded from the ENCODE consortium. Heatmaps on the right show mean nucleotide composition across the top 50 k-mers as ranked by *in vitro* method, by PEKA using Clippy or PEKA using narrowPeaks. (B) Boxplots show differences in mean nt composition of top 50 k-mers as ranked by PEKA using Clippy or narrowPeaks, for 215 eCLIP datasets, which had sufficient tXn coverage in narrowPeaks for PEKA analysis. (C) K-mer logo shows sequence features of 20 most enriched 5-mers for HNRNPC as ranked by RBNS. (D) Mean coverage of 20 most enriched 5-mers for HNRNPC (in RBNS) around crosslink sites for HNRNPC eCLIP and SMInput. The top plot shows HNRNPC eCLIP and SMInput in HepG2 cells and the bottom plot shows HNRNPC eCLIP and SMInput K562 cell line. (E) Number of tXn detected by PEKA in the "protein-coding gene" region when using Clippy or narrowPeaks (paired t-test, $p < 0.0001$) for 57 eCLIP datasets shown in panel A.

To evaluate whether the varying effect of SMInput on recall could be caused by the background model used in PEKA, we selected ten eCLIP datasets with the highest difference in recall observed between the two sets of peaks and analysed them with STREME (Bailey, 2021). We ran STREME using shuffled peak sequences to construct a background model, thus providing an independent approach to investigate the effects of using SMInput for analysis of eCLIP data. We provided STREME either with sequences from Clippy peaks or narrowPeaks, and then compared the ranking of significantly enriched k-mers retrieved by PEKA or STREME (motifs with $p < 0.05$) in the corresponding *in vitro* data (see Methods for details). In the cases of RBPs where PEKA analysis of narrowPeaks gave better ranking than Clippy peaks, STREME analysis also achieved better ranking with narrowPeaks for all datasets (Figure 3.7A). Furthermore, in the cases of RBPs where PEKA analysis of Clippy peaks gave better ranking than narrowPeaks, STREME analysis also recovered motifs with either higher or similar ranking for Clippy peaks compared to narrowPeaks (Figure 3.7B). This indicates that the lack of general motif improvement from narrowPeaks as compared to Clippy peaks in PEKA is reliable, since the alternative motif analysis approach follows the same trends. We also found that the median number of tXn that overlap with narrowPeaks is around 10-fold lower compared to the Clippy peaks used with PEKA (Figure 3.6E). Thus, we find the SMInput filtering in narrowPeaks ineffective in improving the general specificity of identified motifs; however, it greatly reduces the sensitivity of the analysis.

Taken together, the value of SMInput-based correction depends on the specificity of the studied RBP and its abundance in the SMInput control. We speculate that the performance increase observed for RBPs that bind A-rich motifs occurs because these RBPs are minor contributors to the SMInput data, which is dominated by proteins binding U-rich motifs. Conversely, for target RBPs that are abundant also in the SMInput control—like hnRNPC—the SMInput represents a mixture of foreground and background signal, leading to depletion of the foreground signal from narrowPeaks. This analysis demonstrates that while using SMInput controls can improve motif discovery in some instances, it does not lead to overall improvement compared to using only the intrinsic background in CLIP. Therefore, it is appropriate for PEKA to perform motif analysis by relying on the intrinsic background, without employing SMInput filtering.

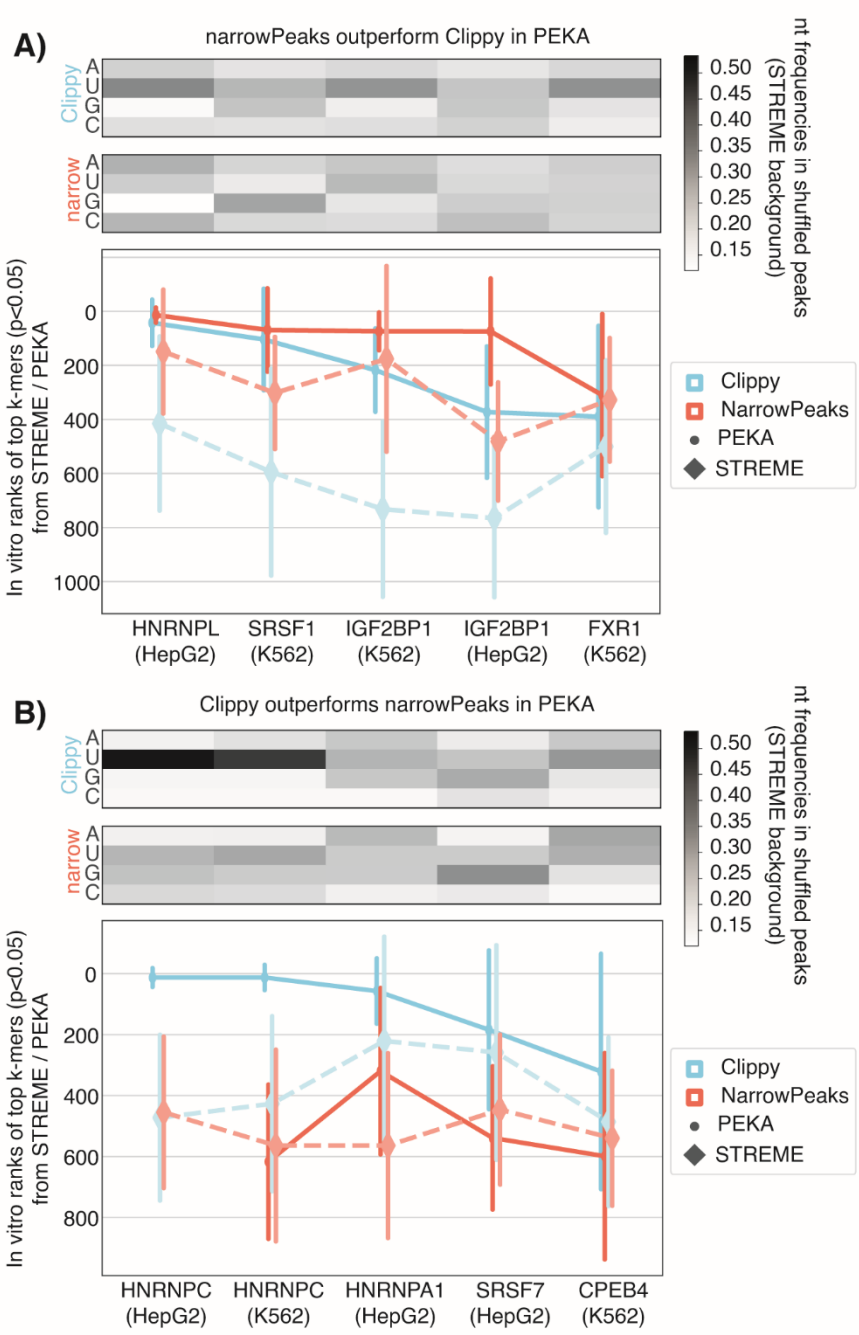


Figure 3.7: Influence of size-matched input controls on motif discovery in PEKA and STREME.

(A, B) Comparison of motif discovery by PEKA and STREME using either Clippy peaks or narrowPeaks, for five eCLIPs where the use of narrowPeaks increased (A) or decreased (B) recall the most, when compared to PEKA run with Clippy peaks. Point plot shows the median ranking of significantly enriched 5-mers ($p < 0.05$) for each combination of motif discovery method (PEKA / STREME) and peaks (Clippy / narrowPeaks) in the corresponding *in vitro* dataset (Methods). Vertical lines represent the standard deviation. Heatmaps show nucleotide frequencies in shuffled peak sequences that were used as background in STREME with Clippy peaks or narrowPeaks.

3.4 Evaluation of Enriched Motifs Across Diverse eCLIP Datasets

We observed that the motifs enriched in different CLIP datasets are influenced by certain sequence biases, notably the U and G-rich preference in eCLIP, and the U-rich preference in PAR-CLIP. We asked whether these biases lead to common motif enrichment patterns detectable in individual CLIP datasets, and how prevalent they might be. To study this, we compared enriched motifs across the entire eCLIP dataset (n = 223 experiments), encompassing 150 diverse RBPs.

We used hierarchical clustering to order the eCLIP data, based on the similarity of their k-mer enrichment (represented by 5-mer ranks), and distribution of thresholded crosslinks in genomic regions comprising protein-coding genes (5′-UTR, CDS, 3′-UTR, and introns). To combine the motif enrichment and the regional binding preferences for clustering, we first calculated two similarity matrices, using the cosine metric: one for the regional information and one for 5-mer ranks. Then, we added the similarity matrices with weights 0.3 and 0.7, respectively, to obtain one combined matrix. Finally, we transformed the combined cosine similarities into cosine distances and used these to perform hierarchical clustering with the *scipy.hierarchy* v1.7.3 module; the linkage was calculated by the Farthest Point Algorithm, also known as a Voor Hees Algorithm (i.e., method = 'complete'). This identified groups of eCLIPs with similar motif enrichment signatures and regional binding preferences (Figure 3.8). For ease of interpretation, we also clustered the k-mers based on their sequence similarity, and arranged these clusters based on their ranking across eCLIP data (see Methods); this identified groups of k-mers with similar sequences, which are enriched for specific groups of eCLIP datasets.

The first visually apparent feature of this analysis is that regional crosslinking preferences are accompanied by trends towards certain motif preferences. For instance, if crosslinks are primarily in 5′-UTR and CDS, the largest cluster of data shows enrichment in GC-rich motifs. In case of datasets with primary crosslinking in introns, motif enrichment is dominated by three clusters: two large clusters dominated by G-rich motifs, and a smaller cluster dominated by U-rich motifs. Datasets that crosslink primarily to 3′-UTRs also show enrichment of U-rich or UA-rich motifs (see the blue and yellow cluster, respectively). These clusters likely include binders of the AU-rich elements that are common regulators of RNA stability in 3′-UTRs (C. Y. Chen & Shyu, 1995). Moreover, both CDS and intronic datasets are often enriched in C/G-rich motifs. This analysis demonstrates that the common motif preferences in eCLIP data are closely linked to the regional crosslinking profiles within protein-coding genes.

As expected, eCLIP datasets within the largest clusters share highly similar motif preferences and regional profiles. We noticed that RBPs falling within these large clusters generally lack orthogonal *in vitro* binding data (indicated by the absence of recall) and are often poorly detected in the mRNA interactome proteomics (enhanced RNA interactome capture, i.e., eRIC) (Perez-Perri et al., 2018), and their top 50 ranked k-mers are often G-, U- or GC-rich. It has been reported previously that such k-mers tend to be overrepresented in eCLIP compared to RBNS (Van Nostrand et al., 2020), but the scale of their presence across eCLIP data was not yet examined. Strikingly, several G-rich k-mers were enriched among the top 50 k-mers in more than 50% of all eCLIP datasets (Figure 3.8).

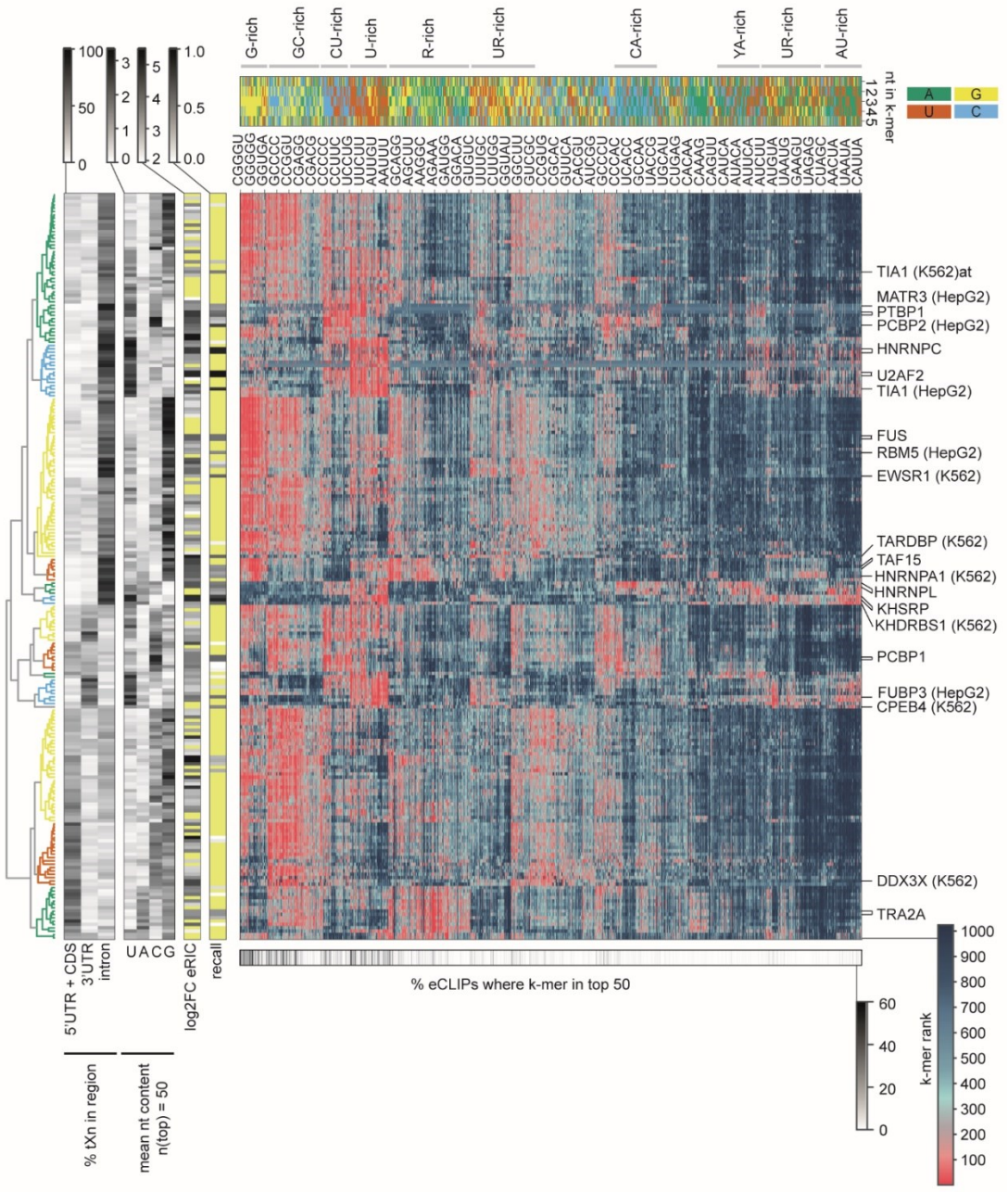


Figure 3.8: Heatmap of all 5-mers for all available eCLIP datasets.

eCLIP datasets are hierarchically clustered by their rank order of 5-mers and regional distribution of thresholded crosslinks to visualise binding preferences across groups of proteins. Fourteen primary clusters were identified, which are colour coded in the dendrogram on the left. Additional heatmaps on the left represent (from left to right): the regional distribution of thresholded crosslink sites; mean nucleotide composition of top 50 identified k-mers; enrichment of the protein in eRIC (Perez-Perri et al., 2018); and the overlap with orthogonal *in vitro* methods (recall). RBPs marked on the right have a recall value > 0.5. A heatmap above the main heatmap represents the nucleotide sequence of each 5-mer and a grayscale heatmap below the main heatmap shows a percentage of eCLIPs, where k-mer ranked among the top 50. Every 20-th 5-mer is labelled on the main heatmap. Yellow fields in grayscale heatmaps indicate missing values.

3.5 Specificity and Sensitivity of CLIP Data in the Context of Sequence and Structural Features of RBPs

In the following sections, we investigated how the sequence specificity of CLIP datasets, measured by PEKA, relates to indicators of data sensitivity and structural features of RBPs. We evaluated eCLIP and PAR-CLIP datasets, using clustering analysis.

3.5.1 eCLIP

To understand how the enrichment of common motifs in eCLIP datasets relates to various features of RBPs, we measured the inter-data similarity of each eCLIP dataset with a similarity score. Similarity score measures how similar are the top 50 k-mers of a particular eCLIP dataset compared to all other datasets. It is calculated as a mean of the pairwise overlap ratios on the top 50 k-mers for all eCLIP datasets. A similarity score of 0 would indicate that the top 50 of k-mers in a certain dataset were not ranked among the top 50 in any other dataset. In contrast, higher values of similarity score indicate that top motifs of a specific dataset overlap with top motifs in many other datasets.

We divided the eCLIPs according to whether they had available *in vitro* data, and then clustered each group based either on a combination of similarity score and recall, or just on the similarity score where no *in vitro* data was available, obtaining 7 clusters (Figure 3.9A, Methods). We visualised the similarity scores and recall for all eCLIP datasets, in a heatmap, together with the number of canonical RNA binding domains—defined as the RRM and the KH domain, and various metrics indicative of data sensitivity: the number of thresholded crosslink sites (*n* tXn), the enrichment of RBP in the RNA interactome capture data (eRIC), and the mean PEKA score of top 50 k-mers (Figure 3.9A).

This visualisation revealed major differences between the group of proteins for which *in vitro* data are available (groups 1–4, i.e., "*in vitro* set") and the group for which no such data are available (groups 5–7, i.e., "eCLIP-only set"). The *in vitro* set tends to have high eRIC values, high number of thresholded crosslinks, and high mean PEKA scores across the top 50 ranked k-mers, whereas the eCLIP-only set tends to have low eRIC values, lower number of tXn, and low mean PEKA scores, indicating that the proteins in the eCLIP-only set do not crosslink well and have low motif enrichment, respectively. The great majority of proteins in the *in vitro* set contain a KH and RRM domain and a low-complexity sequence (Figure 3.9B), which are the common characteristics of RBPs (Balcerak et al., 2019). Conversely, proteins in the eCLIP-only set rarely contain KH or RRM domain. Higher prevalence of canonical RNA-binding domains in the *in vitro* set is not surprising, as the great majority of RBNS and RNAC data are in proteins which contain such domains (Dominguez et al., 2018; Ray et al., 2013). Interestingly, the *in vitro* set contains only a small group 4 with a high similarity score, while the eCLIP-only set contains a large group 7 with a high similarity score, accounting for a third (33%) of all eCLIP experiments. Taken together, proteins lacking orthogonal *in vitro* data generally have different features from the rest, and their eCLIP data tends to have lower inter-data specificity (high similarity index) and motif enrichment (low mean PEKA score, eCLIP-only set). This indicates that cross-validation of eCLIP with *in vitro* data cannot be extrapolated to warrant the specificity of eCLIP data without available *in vitro* data, which must be considered when performing meta-analyses on the whole set of eCLIP data.

The most reliable eCLIP experiments are expected to be in group 2, which includes ~5% of datasets with unique k-mer signatures and high agreement with corresponding RBNS or RNAC data, as indicated by their low similarity index and high recall, respectively. This group of RBPs generally ranked highest in the eRIC experiments,

indicating that they crosslink efficiently with RNA, which is consistent with the high number of thresholded crosslinks identified by PEKA. These RBPs contain a median of 3 RRM or KH domains (Figure 3.9B). Thus, the canonical RBPs that crosslink well and contain many RNA-binding domains tend to yield specific and reliable eCLIP datasets. In addition to the high cross-validation, RBPs in group 2 have the highest mean PEKA scores across the top 50 ranked k-mers, implying that the coverage of top k-mers around tXn is much higher than around oXn. In other words, binding affinity of these RBPs is strongly sequence-dependent, requiring the presence of one or more high-affinity binding motifs.

The least reliable eCLIP experiments are expected within group 7, containing ~33% of eCLIP datasets with high inter-data similarity, which lack orthogonal *in vitro* data (Figure 3.9A). ~39% of datasets in group 7 are undetected in eRIC experiments, which indicates that they crosslink poorly or do not crosslink at all to RNA; however, it is possible that the lack of eRIC data for some of these proteins could be due to discrepancies in their expression between HepG2 and K562 cell lines, in which eCLIPs were performed, and the Jurkat cells, in which eRIC was performed. These proteins lack annotated features of RBPs, such as KH or RRM domains (Figure 3.9A,B). This increases the likelihood of the signal being dominated by the most common contaminants of eCLIP experiments, which are likely the abundant and well-crosslinking RBPs. Low-specificity datasets in group 7 predominantly enrich for G-rich motifs, with most crosslinking sites originating from introns (Figure 3.9A). We note that the same features are also prevalent in group 4, which contains eCLIP datasets that have reasonable agreement with *in vitro* data, indicating that many of these RBPs directly interact with the G-rich motifs. Nevertheless, these experiments have high inter-data similarity because very similar motifs are enriched in groups 4 and 7. However, RBPs in group 4 generally have much higher mean PEKA scores than those in group 7, and thus even though both show enrichment of similar motifs, the extent of enrichment is stronger in group 4 (Figure 3.9B).

It is interesting to find many groups with strong regional binding preferences, even though regional preferences had no direct role in clustering the groups (Fig. 6a). For example, groups 1, 2, 4, and 7 all contain predominantly intronic binding. However, G-rich motifs dominate groups 4 and 7, whereas groups 1 and 2 mainly show enrichment of A-rich, C-rich, or U-rich motifs, likely due to its higher data specificity. It is notable that proteins in group 2 contain a median of 3 KH or RRM domains, whereas those in group 4 contain only a median of 1 domain, and do not ever contain a KH domain. Moreover, group 5 commonly shows predominant binding in CDS and 5′-UTR, which tends to be associated with a higher proportion of A-rich motifs. Since A-rich motifs are otherwise rare in eCLIP experiments, their enrichment contributes to the low similarity index of datasets in group 5. We propose two possible explanations why datasets with similar specificity tend to have similar regional binding. First, the signature might reflect similar contaminants; for example, RBPs that bind G-rich motifs in introns might be the common contaminants of datasets from group 7. Second, the link between RBP specificity and regional binding could reflect regional sequence biases.

In addition to the differences in the content of RRM/KH domains, we found that our clusters of proteins also differed in other domains. The proteins in the *in vitro* set very rarely contain domains other than RRM/KH, whereas the proteins in the eCLIP-only set frequently contain helicase domains, TNase-like domain, Tudor domain, and dsRNA-binding domains (DRBM) (Figure 3.9C). These domains are less sequence-specific, which likely contributes to the generally low eRIC scores of these proteins, and the high similarity scores and low PEKA scores of their eCLIP data. Detailed structural information for analysed RBPs is available in the supplementary materials of our paper ((Kuret et al., 2022), see Additional file 10: Table S9).

We also observed differences in the number and types of compositional biases between groups (Figure 3.9D). In the *in vitro* set, group 4 stands out as its proteins generally have higher numbers of IDRs compared to other groups. Interestingly, group 4 generally binds G-rich motifs and has decent recall and high similarity score. Thus, these proteins generally produce specific data, but their IDRs might be prone to formation of condensates or aggregates that can co-purify in immunoprecipitations of other proteins, which could explain high similarity of their data with many other proteins. Finally, we observed differences in the median PEKA score of each group, which is the highest in groups 2 and 4, which have the highest recall. Thus, the extent of motif enrichment (as quantified by PEKA score) is related to the extent of cross-validation with *in vitro* data (Figure 3.9A,B).

Taken together, we observe a prevalence of G-rich motifs in a third of all eCLIP datasets, with the majority of thresholded crosslinks located in the intronic regions. We show that the high inter-data similarity correlates with lower content of canonical RNA-binding domains and lower data sensitivity, as demonstrated by a lower number of thresholded crosslinks, lower enrichment in eRIC and lower PEKA-scores. This shows the value of cross-RBP examination of enriched motifs in the context of diverse metrics to inform on data reliability. Nevertheless, further studies will be needed to distinguish datasets where high similarity of enriched motifs reflects biological relationships between studied RBPs from those where it reflects common contaminating RBPs or other technical biases in the data.

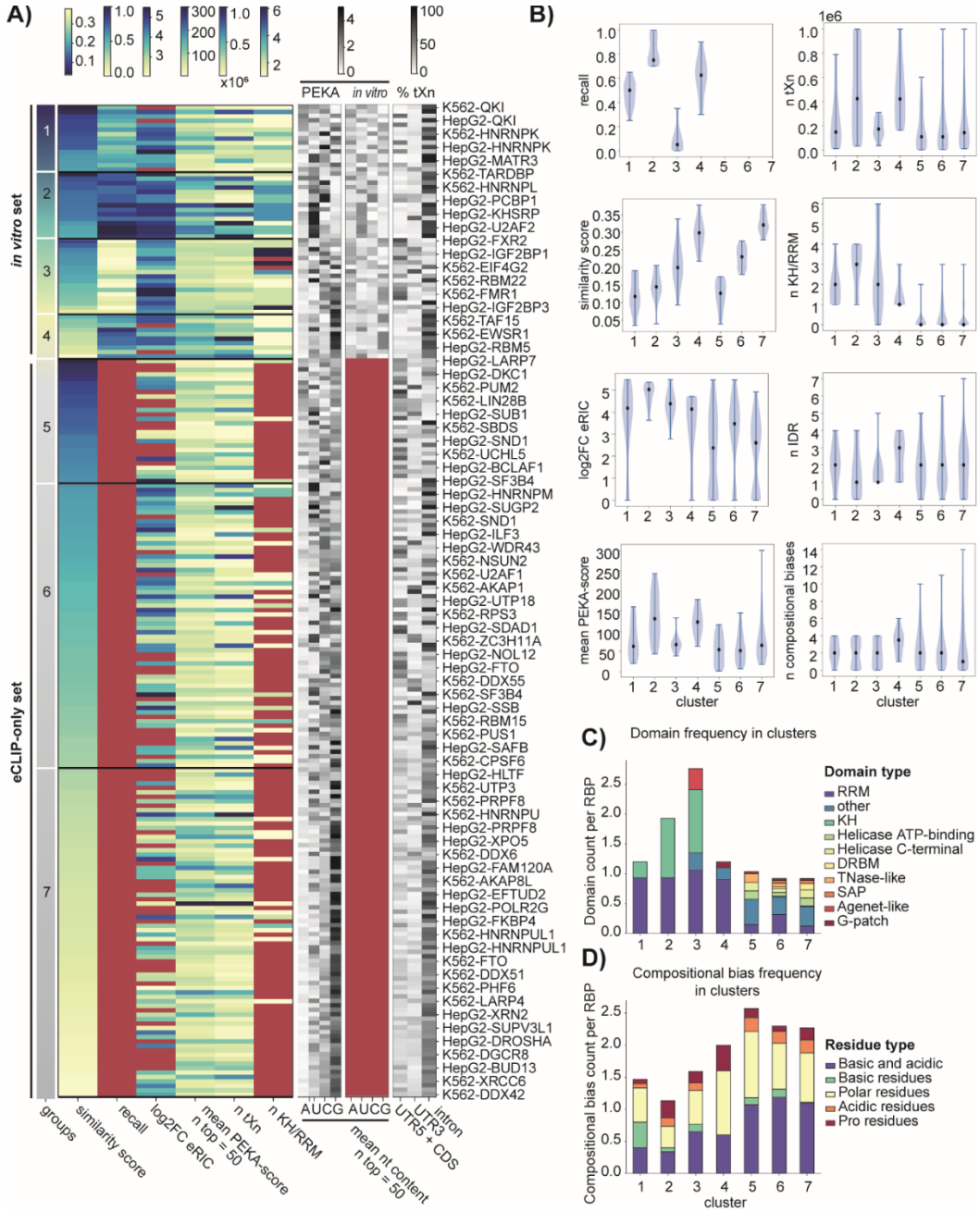


Figure 3.9: Expected specificity of eCLIP datasets with regard to various features.

Continued on next page.

Figure 3.9: *Continued from previous page.*

(A) The left heatmap displays all eCLIP datasets, clustered based on their similarity scores and recall into 7 clusters. For each dataset, the heatmap also shows its enrichment in the mRNA interactome proteomics (log2FC eRIC), the number of KH or RRM in RBP, the average PEKA score across the top 50 ranked k-mers, and the number of tXn in protein-coding region (n tXn). Grayscale heatmaps from left to right show the mean nucleotide content across the top 50 ranked 5-mers (see Methods) for each dataset in PEKA (left) and in vitro data (middle) and % of thresholded crosslinks derived from each transcript region (right). (B) Violin plots show quantitative distribution of features within clusters. In addition to features presented in the heatmap in (A), they also show the number of intrinsically disordered domains (IDRs) and the total number of compositional biases for RBPs within each cluster.
(C, D) Stacked bar plots showing the frequency of a particular domain (C) and compositional biases (D) per RBP within each cluster. Frequency is expressed as a total count of domain or compositional bias per RBP.

3.5.2 PAR-CLIP

To understand the generality of insights obtained from eCLIP meta-analysis, we performed the same analysis for PAR-CLIP experiments, covering 69 diverse RBPs (Figure 3.10). Unlike the eCLIP analyses, we observed no correlation between unique motif enrichments (low similarity index) and high numbers of RRM/KH domains, and instead we observed a slightly reverse trend in PAR-CLIP, with the most unique data obtained for RBPs that lack or have few KH/RRM domains. It is difficult to interpret the relationships between the sequence specificity in PAR-CLIP; as shown in Figure 3.5, PEKA performed worse on the recall metric in PAR-CLIP, compared to eCLIP, and therefore the discovered sequence signatures might not be as reliable. Moreover, evaluated PAR-CLIPs originate from several distinct studies, which may decrease data similarity. Since the purification procedure in most PAR-CLIP experiments employed stringent immunoprecipitation and quality control on SDS-PAGE, it is expected that the extent of contaminating signal is lower in PAR-CLIP, which could explain the lack of correlation between inter-data specificity and data reliability. Furthermore, some differences in enriched motifs in PAR-CLIP compared to *in vitro* data may relate to the unpredictable effects of 4SU on RBP binding, as well as on other aspects of RNA processing (Altieri & Hertel, 2021).

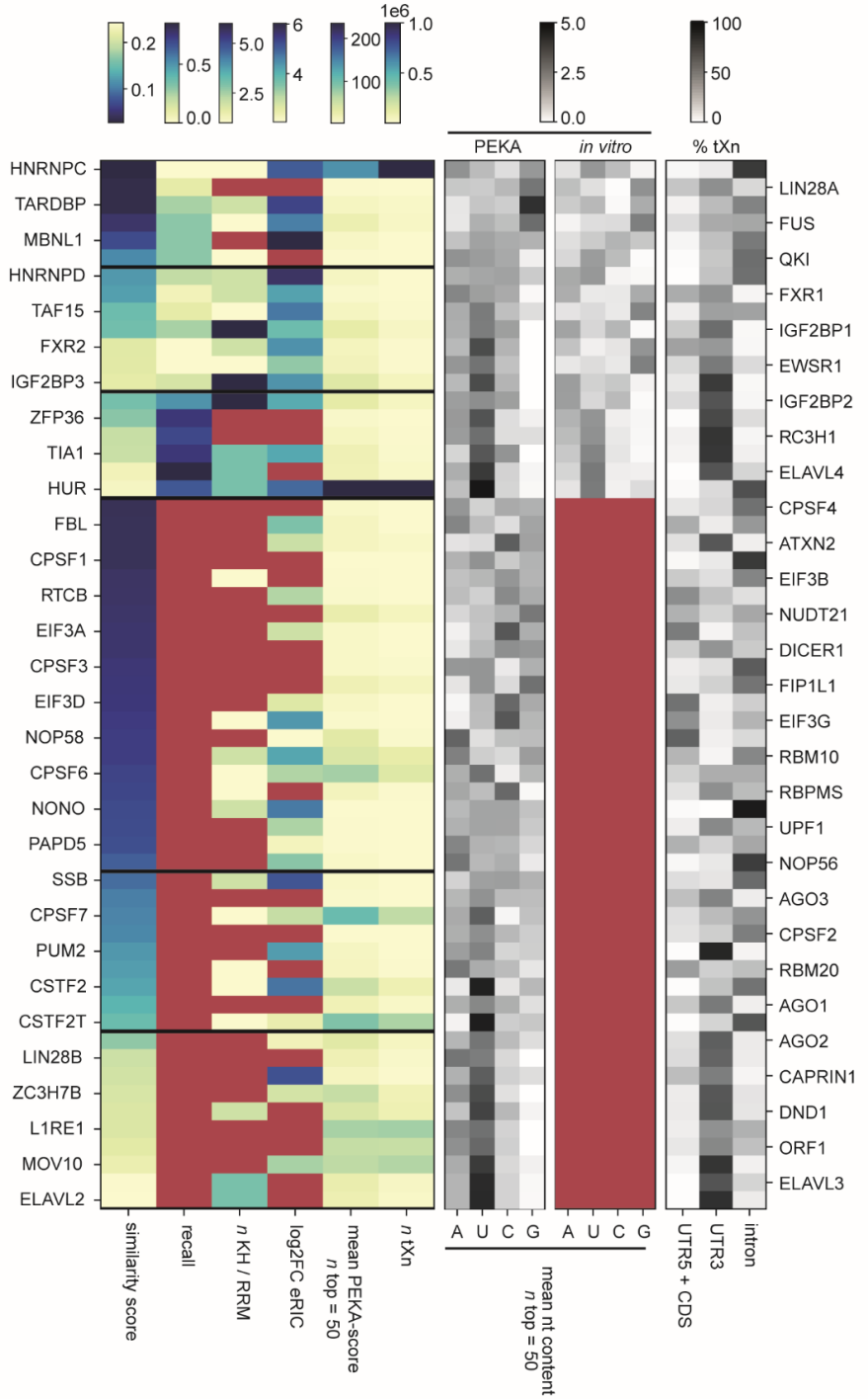


Figure 3.10: Expected specificity of PAR-CLIP datasets with respect to various RBP features.

Heatmap on the left shows 61 PAR-CLIP datasets, clustered based on their similarity scores and recall into 6 clusters. For each dataset the heatmap also shows its enrichment in the mRNA interactome proteomics (log2FC eRIC), the number of KH or RRM in the RBP, the average PEKA-score across the top 50 ranked k-mers, and the number of thresholded crosslinks in the protein-coding gene region. Grayscale heatmaps from left to right show mean nucleotide content across the top 50 ranked 5-mers for each dataset in PEKA (left) and in vitro data (middle) and % of thresholded crosslinks derived from each transcript region (right).

3.6 Summary

In this chapter, we explored the characteristics of motifs enriched at RBP binding sites in CLIP experiments, by either PEKA or mCross, and how their properties relate to the methods' technical biases present at putative crosslink sites, the motifs enriched in *in vitro* data, indicators of data sensitivity, and the structural features of RBPs. Below is a summary of how presented results address the proposed hypotheses.

Hypothesis 1: We can increase the probability of identifying biologically relevant enriched motifs from CLIP data by employing low-count crosslink sites to model background sequences. Our results show that the background model and the enrichment calculation used by PEKA successfully minimises the sequence biases associated with crosslinking events, which reflect the technical biases of the method (Figure 3.3). This is shown for both eCLIP and PAR-CLIP method (Figure 3.3, Figure 3.5), however, PEKA achieves a higher accuracy of motif discovery in eCLIP (Figure 3.5B), indicating that PEKA's background model may be better suited to deep CLIP datasets, in which a sufficient proportion of identified crosslink sites represent technical noise, which allows for effective representation of background. When compared with mCross on eCLIP data, we show that while both PEKA and mCross reduce the U- and G-rich sequence preferences at crosslinks, the mCross method shows a strong bias against the detection of A/U-rich motifs from eCLIP data, which is not the case in PEKA (Figure 3.3C). Therefore, PEKA can perform more accurately on the task of motif discovery, for RBPs that bind to A/U-rich motifs, as demonstrated in Figure 2.4. Conversely, our analysis shows that mCross performs better on RBPs that do not bind U-rich motifs (Figure 2.4). Taken together, the background model in PEKA is equally effective as the one used by the mCross method, suggesting that while using low-count crosslinks to model background is effective, it does not generally outperform the model used by mCross; while mCross worked better than PEKA on some RBPs, PEKA performed better on other RBPs, with respect to their sequence specificity.

Hypothesis 2: Datasets produced by each variant of CLIP method will exhibit some common motif enrichment patterns that can inform on its technical biases. Our research shows that eCLIP experiments exhibit pervasive sequence biases toward uridines and guanosines in enriched motifs for diverse RBPs (Figure 3.8), which are not replicated in iCLIP or PAR-CLIP method (Figure 3.2). These biases correspond to those observed at the crosslink sites in eCLIP, suggesting that the detected patterns of commonly enriched motifs reflect technical biases (Figure 3.3). In PAR-CLIP, we often see the enrichment of U-rich motifs, however, the common motif enrichment patterns, measured by the similarity index, are less pronounced (Figure 3.10); this could be attributed to various reasons, such as PAR-CLIP data being obtained from multiple studies, and a lower level of non-RBP specific background, due to the method's design. So, while our results clearly confirm the presence of common motif enrichment patterns that reflect technical biases in eCLIP, we do not unambiguously confirm similar biases in the collection of PAR-CLIP datasets. Further studies will be required to better understand how technical biases are reflected in enriched motifs in other popular CLIP methods, such as iCLIP and PAR-CLIP, however, this requires large resources of CLIP data with low batch variation, encompassing a diverse set of RBPs, with distinct binding preferences.

Hypothesis 3: The binding specificity of RBPs is affected by the protein's structural features, its IDRs, and its genomic regional binding preferences. Our meta-analysis of eCLIP datasets shows that the linear sequence specificity of RBPs correlates with the content of canonical RNA-binding domains in the protein (Figure 3.9). Conversely, the absence of canonical RBDs and higher content of IDRs and low-complexity regions are

related to lower sequence specificity of RBPs and higher enrichment of common motif patterns, in addition to exhibiting lower data sensitivity. We also show that the enrichment of certain motif groups is characteristic for specific genomic regions—for example the enrichment of AU-rich motifs in 3′-UTRs (Figure 3.8). In PAR-CLIP, we do not observe a relationship between the protein's structure and the inter-data specificity of enriched motifs, but we do see a relationship between a preference for binding to 3′-UTRs and the enrichment of U-rich motifs (Figure 3.10), indicating that regional binding preferences may be reflected in the enriched motifs. Based on our results we conclude that RBPs without canonical RBDs are less sequence-specific than canonical RBPs, and that non-canonical RBPs often produce lower-quality CLIP data. A subsequent study of *in vitro* binding preferences of non-canonical RBPs has since confirmed that most of them lack sequence specificity (Ray et al., 2023). We and Ray et al. propose that non-canonical RBPs may interact with RNA through various mechanisms, such as direct RNA binding that is usually of low specificity, indirect recruitment to RNA via protein co-factors, and stochastic RNA interactions driven by high density of biomolecules, such as within biomolecular condensates. Such modes of interaction would explain our finding that for these RBPs, CLIP data generally contain a lower proportion of high-count crosslink sites and are less likely to be detected in RNA interactome capture experiments.

In addition to addressing the proposed hypotheses, our study showed that the benefit of SMInput filtering for motif discovery is limited to RBPs with a particular sequence specificity—such as the preference towards A-rich motifs—and to RBPs that do not strongly contribute to the signal in the SMInput control. The variation in motif discovery performance between different RBPs shows that what works well for some RBPs, might not work well for others. Therefore, it is important to consider that SMInput filtering decreases the sensitivity of data and is thus only valuable when it strongly enhances data specificity, but it can for some RBPs also be detrimental to data specificity.

Chapter 4

Specificity of LIN28A-Mediated mRNA Decay in Early Embryonic Development

In this chapter, we extend the approach of comparative motif analysis to ask how modifications in IDRs of LIN28A modulate its function. LIN28A is a key regulator of developmental timing, controlling cell proliferation, stemness, and growth through distinct mechanisms. Its involvement in these diverse cellular processes is underlined by its capability to interact with various biological molecules, such as miRNA, mRNA, DNA, and protein, through its unique molecular structure, comprising a CSD, two zinc-finger motifs and N- and C-terminal disordered regions (Figure 4.1A,B). Our work focuses on understanding LIN28A's role in early embryonic development, specifically during the naive-to-primed cell fate transition, which occurs within 24 hours of blastocyst implantation in mice and coincides with the polarisation of cells and their assembly into a rosette structure. During this transition, LIN28A gets phosphorylated in its C-terminal IDR (Tsanov et al., 2017) (Figure 4.1A,B) by the MEK/ERK signalling cascade, which is crucial for commitment to primed pluripotency (Neagu et al., 2020). Upon MEK/ERK activation, the levels of naïve regulon mRNAs decrease (Vega-Sendino et al., 2021), which is mainly ascribed to enhancer-mediated regulation (Hamilton et al., 2019; Respuela et al., 2016; Vega-Sendino et al., 2021). However, this transcriptional mechanism is hard to reconcile with the rapid and selective mRNA clearance observed by a recent study (Modic et al., 2021).

A study by Tsanov et al. directly linked MEK/ERK-induced phosphorylation of LIN28A in its intrinsically disordered region at residue S200 in humans and mice (Figure 4.1A,B) with the promotion of the naïve-to-primed transition. Moreover, the study showed that upon phosphorylation, LIN28A interacted more with the mRNA targets rather than miRNAs, which are more extensively studied as the mechanism of developmental timing regulation by LIN28A (Tsanov et al., 2017). The residue S200 (in mice) is conserved across mammalian species; however, potential phosphosites in this region are also present in zebrafish and chicken LIN28A (Figure 4.1A), suggesting that the phosphorylation in this region of the IDR may be relevant for the function of LIN28A in other cellular processes that are not specific to mammalian development.

To explore these questions, we apply bioinformatics to analyse gene expression in multiple cell states—measured with the 3′end RNA-sequencing experiments—and binding patterns of LIN28A and PABPC proteins, measured by iCLIP experiments. 3′-end sequencing experiments and iCLIP experiments were designed to recapitulate the cell states

of naïve-to-primed transition, while varying LIN28A expression and phosphorylation. To simulate naïve-to-primed transition, the pluripotent stem cells were first cultured in the medium that inhibits MEK/ERK signalling and promotes WNT signalling and cell differentiation (the '2i/LIF medium'); this maintained the cells in the state of naïve pluripotency. To activate the cell-fate transition, the naïve cells were then transferred to the 'FGF2 medium', which activated the MEK/ERK signalling cascade (see Methods for details). These cell culture conditions enabled for precise timing of MEK/ERK activation and allowed us to study RNA decay in the context of LIN28A phosphorylation.

Using these approaches, we revealed a novel regulatory mechanism of LIN28A, in which the phosphorylated LIN28A relocates to AU-rich 3′-UTR termini bound by cytoplasmic poly(A)-binding proteins, which, in turn, triggers their decay, and enables cell differentiation. We also found that the selectivity of mRNA decay is mediated by RNA sequence, as well as greater accumulation of LIN28A and poly(A)-binding proteins to these regulatory regions. Specifically, mRNAs targeted for decay exhibited higher multivalency of AU-rich motifs and a greater accumulation of both LIN28A and poly(A)-binding proteins at their terminal regions. With this research, we hope to shed light on the crucial role of IDRs in modulating the dynamic RBP functions in development.

4.1 Recent Work that Led to This Study

A study by Modic et al. trained a machine learning classifier to predict developmental mRNA decay and, through feature importance analysis, identified LIN28A and PABPC1 binding to the 3′-UTRs of mRNAs to be predictive of their stability in development (Modic et al., 2021). Moreover, this study found that the expression of LIN28A in conjunction with active MEK/ERK signalling is essential for the coordination of pluripotency progression and the clearance of naïve mRNAs, which preceded the effects of LIN28A on let-7 miRNA processing, indicating that LIN28A acts directly on the mRNAs to mediate their destabilisation (Modic et al., 2021). Conversely, in the absence or enforced nuclear localisation of LIN28A, cell differentiation was delayed, leading to the formation of multiple rosette structures and embryonic lumens (Modic et al., 2021).

4.2 LIN28A Phosphorylation Promotes its Interactions with 3′-UTR

The list of naïve pluripotency genes, which are rapidly downregulated in naïve-to-primed transition, was defined as in the paper by (Modic et al., 2021), on the basis of data published by (P. Yang et al., 2019). For a list of naïve genes analysed in this study see Table 4.1. To understand whether the phosphorylation of LIN28A triggers the decay of naïve regulon mRNAs, we conducted a differential gene expression analysis, comparing the effects of WT LIN28A (LIN28A-WT) and phosphonull LIN28A mutant (LIN28A-S200A) upon induction of MEK/ERK signalling. For this, we analysed the 3′-end sequencing data produced in LIN28A knock-out mouse embryonic cells in three conditions: without the induction of LIN28A-WT transgene expression, i.e., the KO condition; upon induced expression of LIN28A-WT transgene; and upon induced expression of LIN28A-S200A transgene. To decrease the likelihood of off-target effects, and to mimic the biological system as closely as possible, the transgene expression was matched to physiological expression level. In all three conditions we stimulated MEK/ERK activity with the FGF2 medium. We quantified gene expression levels in each condition 6 hours after the switch to MEK/ERK signalling and/or simultaneous induction of LIN28A transgene expression

and compared gene expression between cells expressing the LIN28A-WT transgene vs KO, and cells expressing the LIN28A-S200A transgene vs KO (Figure 4.1C, see Methods for detailed description of 3′-end sequencing experiments and their analysis). This showed that when MEK/ERK is active, the expression of LIN28A-WT leads to significant changes in expression levels (Figure 4.1C,D): downregulating ~1750 genes—including the genes of naïve regulon—and upregulating ~2000 genes. Conversely, the expression of LIN28A-S200A led only to small changes in mRNA levels when compared to KO—66 downregulated and ~250 upregulated genes. In subsequent analyses, we aimed to elucidate different factors that affect selective targeting of mRNAs for degradation by pLIN28A. For this purpose, we defined three groups of genes that are either downregulated (DOWN), unchanged in expression (Control) or upregulated (UP) by phosphorylated LIN28A, as compared to LIN28A KO (Figure 4.1C, left). We alert the reader that we will refer to these three groups of genes throughout the following sections and figures.

To determine if the phosphorylation of LIN28A at S200 generally changes its binding along transcripts, we analysed the proportions of crosslink counts along the transcript's exons for LIN28A-WT in the 2iLIF treated condition, as well as LIN28A-S200A and LIN28A-WT in the FGF2 treated condition (Figure 4.1E, see Methods for details). Our analysis revealed that upon LIN28A phosphorylation, the density of crosslink sites in the 3′-UTR exons increases by approximately 20 % (per kb, relative to other exonic regions), compared to conditions where LIN28A is not phosphorylated. When comparing the exon-binding ratios between UP, Control, and DOWN transcripts, we observed the relative increase in 3′-UTR binding upon LIN28A phosphorylation in all transcript groups, regardless of their stability in the naïve-to-primed transition. Nevertheless, downregulated transcripts showed a slightly greater density of pLIN28A in 3′-UTRs than Control and UP transcripts, suggesting that a higher level of pLIN28A binding to 3′-UTRs is inversely correlated with transcript stability.

Intriguingly, the binding densities between exon regions are very similar between the two unphosphorylated LIN28A states—LIN28A-WT in 2iLIF treated cells and LIN28A-S200A mutant in the FGF2 treated cells—indicating that the increase in the 3′-UTR binding observed for pLIN28A is likely a direct result of LIN28A phosphorylation, rather than other confounding effects that may arise from the activation of MEK/ERK signalling pathway. Together these analyses demonstrate that the phosphorylation of LIN28A at the residue S200 by MEK/ERK signalling increases its interaction with the 3′-UTRs of mRNAs and is required for inducing decay of naïve mRNA regulon.

Table 4.1: A list of naïve genes analysed in this study.

Index	**Gene Name**	**ENSEMBL Gene ID (stable)**
1	ESRRB	ENSMUSG00000021255
2	FOXD3	ENSMUSG00000067261
3	KDM3A	ENSMUSG00000053470
4	KDM3B	ENSMUSG00000038773
5	KLF2	ENSMUSG00000055148
6	KLF4	ENSMUSG00000003032
7	KLF5	ENSMUSG00000005148
8	KLF9	ENSMUSG00000033863
9	NANOG	ENSMUSG00000012396
10	NR5A2	ENSMUSG00000026398
11	PRDM14	ENSMUSG00000042414
12	SALL1	ENSMUSG00000031665
13	SOX2	ENSMUSG00000074637
14	TBX3	ENSMUSG00000018604
15	TCF7L1	ENSMUSG00000055799
16	TCL1	ENSMUSG00000041359
17	TET2	ENSMUSG00000040943
18	TFCP2	ENSMUSG00000009733
19	TFCP2L1	ENSMUSG00000026380
20	ZFP281	ENSMUSG00000041483
21	ZFP42	ENSMUSG00000051176
22	ZIC2	ENSMUSG00000061524
23	ZIC3	ENSMUSG00000067860

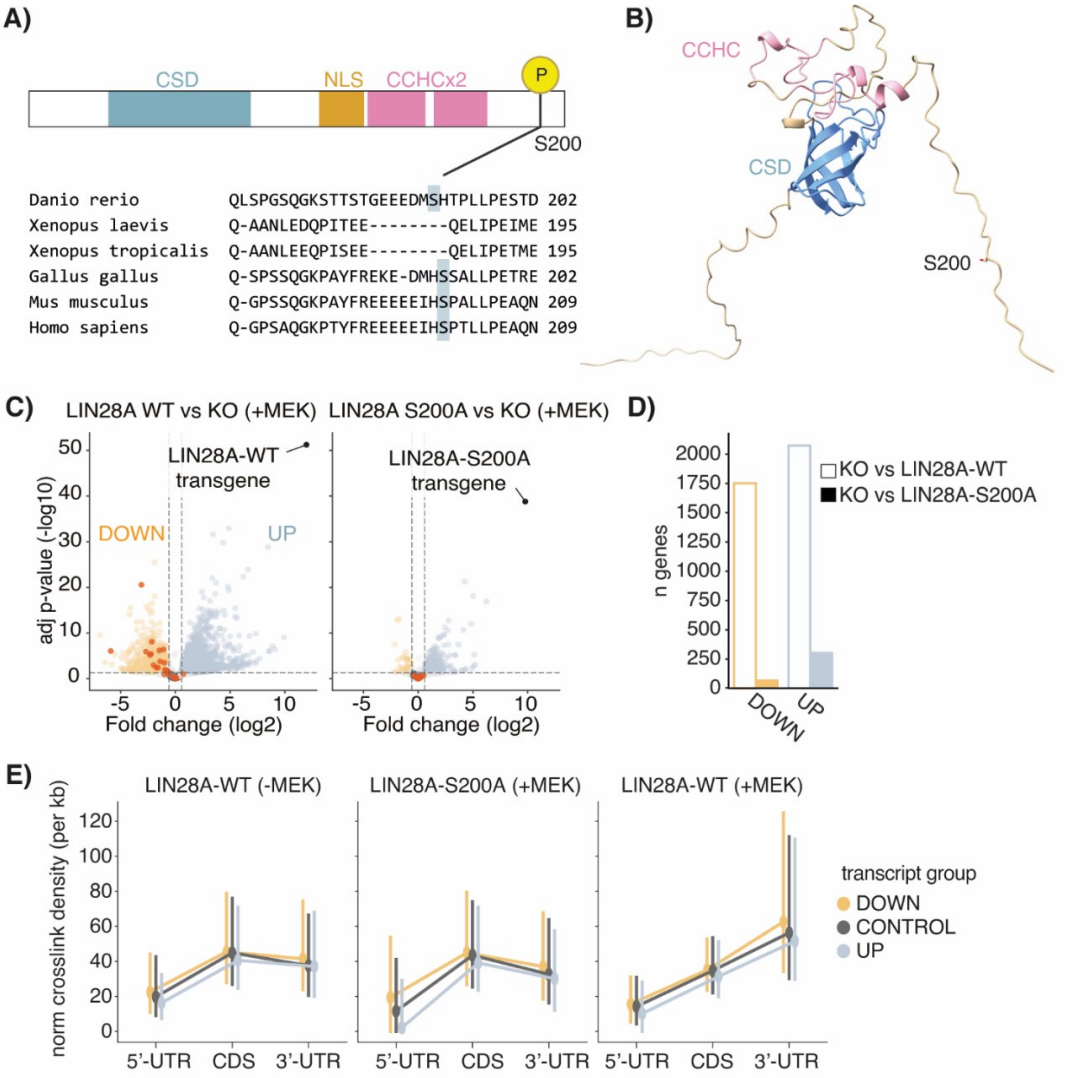


Figure 4.1: Structure of LIN28A and the effects of S200 phosphorylation on differential gene expression.

Continued on next page.

Figure 4.1: *Continued from previous page.*

(A) Schematic representation of LIN28A structure. CSD stands for cold-shock-domain; NLS stands for nuclear localisation signal; CCHCx2 stands for two CCHC zinc-fingers. Below, the sequence conservation of the S200 phosphorylation site (in mice and human), is shown. (B) AlphaFold prediction for mouse LIN28A (Uniprot ID: Q8K3Y3) shows the structure of CSD and zinc-finger domains, as well as the N- and C-terminal disordered regions. The residue S200 is located in the C-terminal disordered region. (C) Differential expression for protein-coding genes when comparing 3′end RNA sequencing data for LIN28A-WT (left) or LIN28A-S200A transgene expression (right) to KO, in the FGF2-treated condition. Differentially expressed gene groups—marked DOWN and UP—were determined by applying a criterion for adjusted p-value <0.05 (α=0.1) and fold-change $\geq$1.5. Control group of genes was defined by applying a criterion for adjusted p-value $\geq$0.05 and |log2 fold-change| $<$0.5. Genes of naïve regulon are marked in red. (D) The number of up- or downregulated genes when comparing 3′end RNA sequencing data for LIN28A-WT (white) or LIN28A-S200A transgene expression (coloured) to KO, in the FGF2-treated condition. (E) Relative quantification of iCLIP signal in transcript's exons (5′-UTR, CDS, and 3′-UTRs) for: LIN28A-WT in the 2iLIF treated condition (-MEK, left); LIN28A-S200A (middle) and LIN28A-WT (right) in the FGF2 treated condition (+MEK). For each transcript we calculated the % of cDNA counts within its exonic regions and normalised the obtained value by region's length to get crosslink density (expressed as % crosslinks per kb). Each pointplot shows the median iCLIP signal in exonic regions across three groups of transcripts—DOWN, Control, and UP; the error bars represent the interquartile ranges.

4.3 pLIN28A Relocates to PABP-Bound Multivalent A/U-Rich 3′-UTR Termini to Promote Selective mRNA Decay

Building on our previous finding that the phosphorylation of LIN28A promotes its interactions with the 3′-UTRs, we wanted to understand whether the accumulation of pLIN28A in these regions is related to specific binding motifs. We used PEKA to identify sequence motifs enriched around LIN28A crosslinks, located within 3′-UTRs, in iCLIP experiments targeting LIN28A-WT and LIN28A-S200, in the FGF2-treated cells, and LIN28A-WT in the 2iLIF-treated cells. We collated significantly enriched k-mers ($p<0.05$) from all evaluated experiments and clustered them, based on their sequence similarity and iCLIP enrichment to obtain three distinct groups of k-mers (Figure 4.2A). Among these were k-mers that corresponded to canonical zinc-finger-bound (WGG, n=24) (Figure 4.2B) or CSD-bound motifs (GAU, n=14) (Figure 4.2C), as well as a group of auxiliary AU-rich motifs (n=55) (Figure 4.2D). In contrast to the WGG and GAU motifs, the AU-rich motifs were not frequently reported to be targeted by LIN28A in the literature and were not detected by *in vitro* RNAcompete method, which only reported the GGAG motif (Ray et al., 2013). However, they were found enriched at mRNA binding sites of LIN28B, detected by PAR-CLIP experiments in HEK293 cells (Mukherjee et al., 2019), as well as in eCLIP experiments analysed in our study (Kuret et al., 2022); see Figure 2.2A for LIN28B eCLIP in K562 cells. We noticed that in the crystal structures of LIN28A bound to the let7 pre-miRNA, CSD forms pi-stacking interactions and hydrogen bonds with the AUU stretch (Figure 4.2D) (Nam et al., 2011). Moreover, contacts of CSD with AUU or UUU motifs are observed in 3 additional structures of LIN28A from 3 organisms (Heinemann & Roske,

2021), indicating that CSD is the likely source of interaction with the auxiliary AU-rich motifs. We visualised the coverage of each group around crosslink sites within 3′-UTRs and observed that LIN28A phosphorylation strongly increases its binding to AU-rich motifs, while concomitantly decreases its binding to the canonical ZnF-bound WGG motifs (Figure 4.2B,D).

We next aimed to decipher whether the changes in LIN28A binding affect its selective targeting of mRNAs for decay. For this, we first examined the density of the iCLIP signal across the three types of exons—5′-UTR, CDS, and 3′-UTR—for iCLIPs of LIN28A-WT before and after the activation of MEK/ERK signalling (Figure 4.3A, see Methods for details). This revealed a striking re-arrangement of LIN28A binding in 3′-UTRs: prior to phosphorylation, LIN28A is dispersed along the 3′-UTR, whereas after phosphorylation, it condenses at 3′-UTR starts, i.e., downstream of the STOP codon, and at the 3′-UTR termini. A less pronounced reorganisation of LIN28A binding also occurs in the CDS, with pLIN28A signal increasing at the CDS bounds. In 3′-UTRs, we observed an interesting difference in positioning of pLIN28A between UP and DOWN transcripts; while UP genes exhibit approximately equal peaks at the 3′-UTR start and end, the DOWN genes show a shift in binding toward the 3′-UTR termini.

Since we were interested primarily in LIN28A's regulation of mRNA decay, we examined LIN28A binding at 3′-UTR termini in more detail, by quantifying absolute iCLIP signal and normalising it to transcript expression and library-size, to enable comparison between different iCLIP samples and transcript groups. This revealed that upon phosphorylation, the binding of LIN28A increases approximately 50nt upstream of poly(A) signal (PAS), in both DOWN and Control or UP transcripts (Figure 4.3B). This region coincides with a high incidence of AU-rich motifs that were previously found to be enriched downstream of pLIN28A crosslink sites (Figure 4.3D). Although pLIN28A binding upstream of PAS occurs in both DOWN and Control/UP transcripts, the density of pLIN28A binding is ~2.5-fold greater in DOWN transcripts, which also exhibit a higher density of AU-rich motifs in this region (Figure 4.3C). In contrast, no significant difference in motif density was observed between transcript groups for canonical LIN28A binding motifs—the WGG and GAU motifs (Figure 4.3C).

To corroborate these observations, we quantified motif content in the 100nt region upstream of PAS by counting the representative trimers for motif groups defined by PEKA (Figure 4.3D, Figure 4.2A): the WUU-motif group was represented with the AUU and UUU trimers; the CSD-binding motif was represented with the GAU trimer; and the zinc-finger binding motif was represented by AGG and UGG trimers. Notably, the valency of WUU trimers in this region was higher in the DOWN transcripts by a factor of 1.6 or 2.3 when compared to Control or UP transcripts, respectively (Figure 4.3D). In contrast, the canonical WGG and GAU trimers known to be recognised by the zinc-finger and CSD domains of LIN28A have slightly lower incidence in DOWN transcripts as compared to Control or UP transcripts (Figure 4.3D). High content of A/U nucleotides at the 3′-UTR termini, and the increased binding of pLIN28A to these regions, is also evident in naïve-regulon mRNAs (Figure 4.3E). It has been shown that multiple binding motifs may need to be clustered within multivalent RNA regions to enable IDR-mediated regulatory capacity of RBPs (Gueroussov et al., 2017; Hallegger et al., 2021), and indeed we found that the region upstream of PAS is particularly multivalent in the naïve-regulon and DOWN transcripts, with a median of 8 WUU-motif trimers within the 100nt upstream of PAS. This suggests an important role of AU-multivalent regions at the termini of 3′ UTRs in the phosphorylation-induced LIN28A repositioning, and thus in selecting mRNAs for developmental decay.

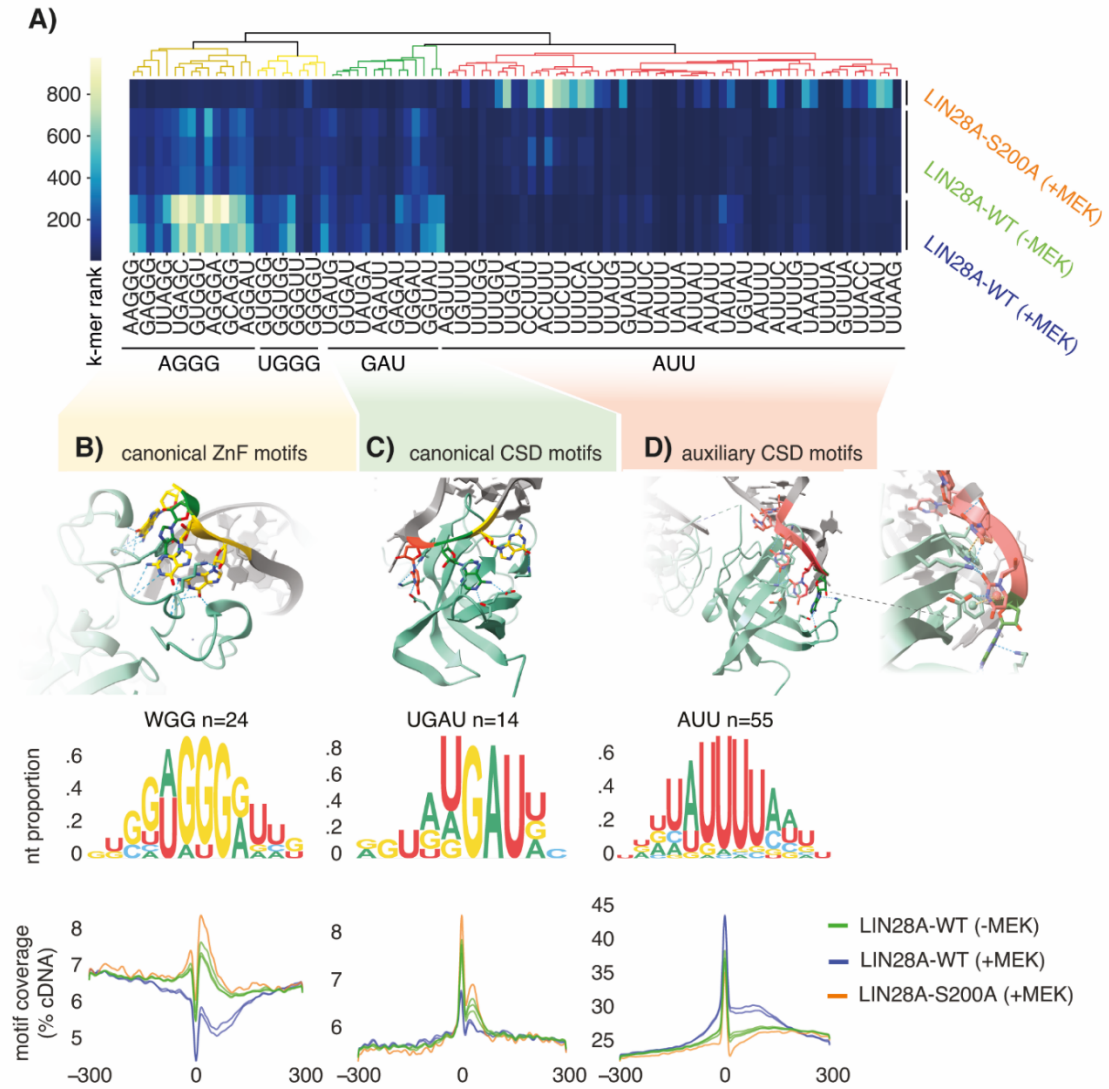


Figure 4.2: Sequence binding preferences of LIN28A in phosphorylated and unphosphorylated states.

(A) Heatmap shows hierarchically clustered rankings of 5mers that were found significantly enriched by PEKA software in any of the LIN28A iCLIP samples. Motif enrichment analysis was performed in the 3′-UTRs. K-mers were ranked based on their enrichment score in PEKA, rank of 1 representing the most enriched 5mer and rank 1024 representing the least enriched 5mer in the dataset. Clustering was performed based on sequence similarity of k-mers and based on their ranking (Methods). (B,C,D) Upper panels show a crystal structure of LIN28A (PDB ID: 3trz (Nam et al., 2011)) in complex with GAGG (B), GAU (C) and AUU (D) RNA motifs through its ZnF (B) and CSD domains (C, D). Middle panels display corresponding meta-motif representations of three motif groups that were identified by PEKA as enriched in 3′-UTRs in LIN28A CLIP data (A). On the bottom, line plots show the distribution of motif group coverage around crosslink sites located in the 3′-UTRs.

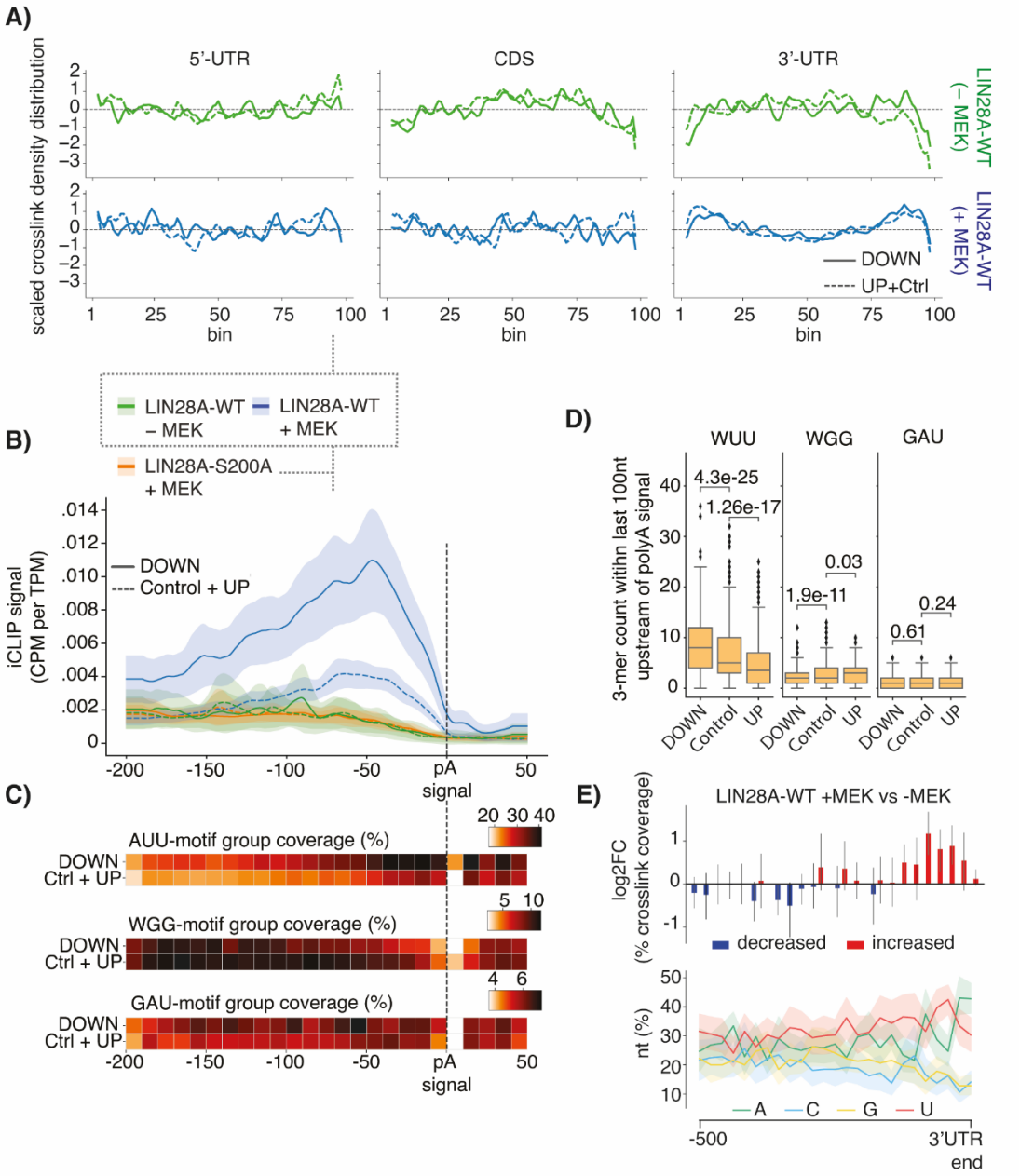


Figure 4.3: pLIN28A converges to AU-rich 3′-UTR termini to trigger selective mRNA decay.

Continued on next page.

Figure 4.3: *Continued from previous page.*

(A) Line plots show the relative distribution of iCLIP signal (Methods) across 5′-UTRs (left), CDS (middle), and 3′-UTRs (right) for iCLIPs of LIN28A-WT in the FGF2-treated condition (+MEK, bottom, blue lines) and in the 2iLIF-treated condition (-MEK, top, green lines). On each plot, the distributions are visualised for two groups of transcripts: those downregulated upon LIN28A phosphorylation (DOWN, solid line) and those upregulated or unchanged upon LIN28A phosphorylation (UP + Control). (B) Line plot indicates the mean of expression-normalised crosslink coverage in the region 200 nts upstream and 50 nts downstream of PAS for iCLIPs of LIN28A-WT and LIN28A-S200A in the FGF2-treated condition (+MEK) and LIN28A-WT in the 2iLIF-treated condition (-MEK). Included 3′-UTRs were filtered for minimum length and expression level as described in Methods. Shaded areas indicate a 95% confidence interval. (C) Heatmaps show the mean percentage of nucleotides covered by AUU- (top), WGG- (middle) and GAU-motif groups (bottom) in DOWN and Control/UP genes (Methods). Mean coverage is calculated across evaluated genes in 10nt bins in a region 200nts upstream and 50nts downstream of PAS. (D) Boxplots show the counts of specific 3-mers in a region 100nts upstream of PAS. The quantified 3-mers are noted on the plot in the IUPAC notation: the WGG encompasses AGG and UGG; the WUU encompasses AUU and UUU. Included genes were filtered as described in Methods to yield 1002 genes in DOWN, 2126 genes in Control, and 778 genes in UP group. Two-sided Mann-Whitney-Wilcoxon test was performed to assess the significance of difference in groups' means; the comparisons and p-values are annotated on the graph. (E) Barplot (top) shows the log2 fold-change in expression-normalised crosslink coverage between iCLIPs of LIN28A-WT in FCL-treated condition (+MEK) vs LIN28A-WT in 2iL-treated condition (-MEK) in genes of naïve regulon (see Methods for details). Fold-changes are shown in the region of 500 nts before the 3′-UTR termini, split into bins of 20 nts. Error-bars represent 95% confidence intervals. Line plot (bottom) shows the mean percentage of nucleotides in each bin, across evaluated genes. The shaded areas represent 95% confidence intervals.

Previous work by Modic et al. showed that besides LIN28A, the binding of PABPC1 to 3′-UTRs was also predictive of mRNA destabilisation during naïve-to-primed transition (Modic et al., 2021). Since LIN28A is not known to catalyse mRNA decay directly, we asked whether it might work in concert with PABP proteins—which play various roles in the regulation of mRNA stability—to mediate this process. Several studies have shown that PABP can promote deadenylation and mRNA decay (He et al., 2023; Webster et al., 2018; Yi et al., 2018). Two separate studies have shown that LIN28A and PABPC1 interact: one reported their direct interaction, while the other reported the interaction as RNA-dependent (Balzer & Moss, 2007; N.-K. Yu et al., 2021). Another study produced a crystal structure of PABPC1's RRM interacting with the cold-shock-domain of a multi-CSD RBP Unr (Hollmann et al., 2023). Based on these findings we hypothesised that LIN28A and PABP work together to regulate mRNA destabilisation in naïve-to-primed transition. To test this hypothesis, we analysed iCLIP data for PABPC1 and PABPC4—which are the only cytoplasmic PABPs in mice—in LIN28A knockout cells with or without induction of LIN28A-WT upon 6h of MEK/ERK activation (PABPC and PABPC4). Additionally, we analysed the iCLIP of PABPC1 in naïve cell state (maintained with 2iLIF medium), where native, unphosphorylated LIN28A was expressed, to evaluate PABPC1 binding prior to the induction of MEK/ERK signalling pathway.

First, we evaluated the density of PABPC binding to mRNA exonic regions in these different conditions (Figure 4.4A) and found that the ratio of binding density between the exons resembles that of phosphorylated LIN28A-WT (Figure 4.1E) with the highest density

in the 3′-UTR regions. Additionally, in all iCLIP experiments, the density of PABPC binding in the 3′-UTR was the highest in transcripts that were downregulated by pLIN28A induced decay. Surprisingly, these patterns were observed for PABPC1 even in naïve cell state, where MEK-signalling was inactive (Figure 4.4A).

To understand the location of PABP binding on the transcripts in greater detail, and to find whether UP/Control and DOWN transcripts exhibit distinct binding patterns, we next investigated the crosslink density profiles (Figure 4.4B). These revealed that both PABPC1 and PABPC4 bound to the same transcript regions—primarily the CDS and the 3′-UTR bounds—irrespective of LIN28A expression and MEK activation. Specifically, this binding pattern was observed for PABPC1 in naïve cell state, where native LIN28A was expressed (Figure 4.4B, green lines); and upon activation of MEK/ERK signalling cascade in LIN28A KO cells, with (Figure 4.4B, red lines) or without (Figure 4.4B, blue lines) the induction of LIN28A-WT transgene expression.

Moreover, we observed that while the same transcript regions are bound in UP/Control and in DOWN transcripts, the ratio of PABPC binding at the CDS and 3′-UTR bounds differs between the two groups of transcripts, when MEK/ERK pathway is activated (Figure 4.4B, red and blue lines). At the stop codon, UP/Control transcripts had a higher binding peak, compared to DOWN transcripts, and the reverse was true for 3′-UTR termini, where DOWN transcripts showed a higher binding peak (Figure 4.4B, red and blue lines). This pattern in the 3′-UTR matches that observed for pLIN28A (Figure 4.3A), indicating that the LIN28A and PABP proteins bind to the same transcript regions, and respect the same shifts in binding between UP / DOWN transcripts. Interestingly, the PABPC1 binding in naïve cell state (Figure 4.4B, green lines) did not show any shifts in binding ratios at CDS bounds but did show increased accumulation towards 3′-UTR termini.

Next, we evaluated the co-localisation of LIN28A and cytoplasmic PABPs on a granular level for 3′-UTRs of three naïve genes: Tfcp2l1, Esrrb and Zfp281 (Appendix A.1-A,D,E). Surprisingly, we observed that after phosphorylation LIN28A remained bound to sites overlapping with PABP peaks but did not retain its binding elsewhere in the 3′-UTR (Appendix A.1-B). Moreover, we observed that the binding of pLIN28A specifically increased, compared to the unphosphorylated LIN28A, at PABP peaks within the terminal 200nt of 3′-UTR (Appendix A.1-D). Analogously to LIN28A, we inspected PABP binding at 3′-UTR termini in more detail (Figure 4.4C) and found that both PABPC1 and PABPC4 also bind to the AU-rich region upstream of PAS. Moreover, the quantification of expression normalised iCLIP signal revealed that when LIN28A is expressed, DOWN transcripts exhibit ~5-fold increase in signal in this region, compared to Control or UP transcripts. Conversely, if LIN28A was not induced the difference between transcript groups was minor.

Previously, we found that DOWN transcripts exhibit a greater density of A/U-rich motifs, compared to Control or UP genes, however, these sequences are not usually associated with PABPs. To evaluate whether the terminal 3′-UTR regions of DOWN genes might be more multivalent also in the polyA motifs, which are known to be bound by PABPs, we quantified the occurrence of the AAA 3-mer in the 100 nt region upstream of PAS (Figure 4.4D). Indeed, we found a linear trend of decreasing AAA content, with DOWN transcripts having the highest valency of AAA and UP transcripts having the lowest valency. The increased valency of AAA in DOWN transcripts could explain the greater accumulation of PABPs, irrespective of LIN28A, in these regions (Figure 4.4C), and with this confer the selectivity of pLIN28A-induced transcript destabilisation.

Together, these results indicate that upon phosphorylation, LIN28A relocates to the regions at 3′-UTR termini that are also bound by cytoplasmic PABPs. PABPs bind to these RNA regions prior to the activation of MEK/ERK pathway and do so independently

of LIN28A expression (Figure 4.4B). The convergence of pLIN28A to these poised PABPC sites in turn leads to greater association of PABPC proteins with the A/U-rich regions upstream of the PAS to promote transcript destabilisation. Together this leads to the destabilisation of naïve mRNAs, which drives cell differentiation from naïve to primed pluripotency (Figure 4.5).

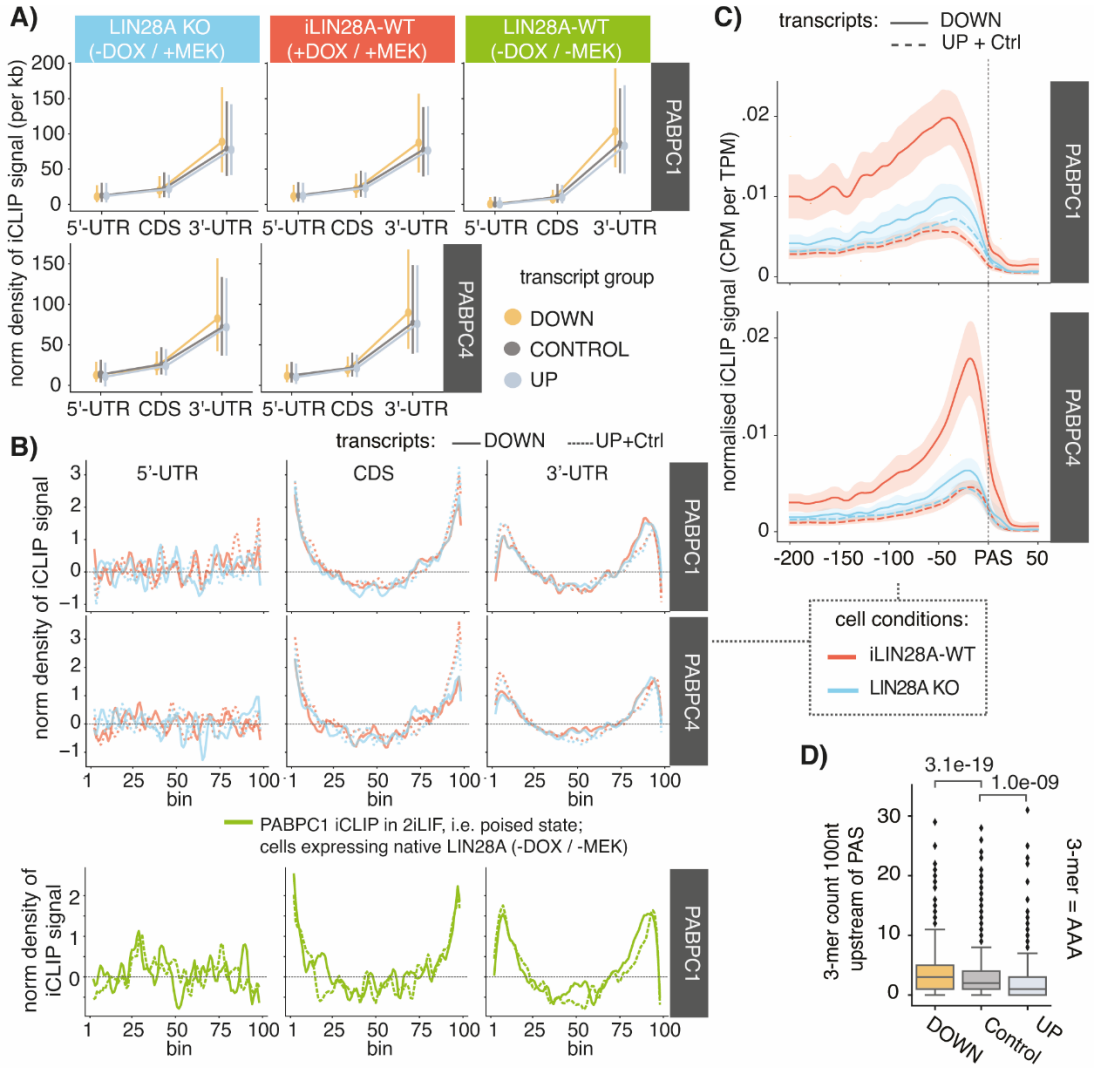


Figure 4.4: Increased PABPC1/4 binding to 3′-UTR termini correlates with pLIN28A-mediated transcript destabilisation.

Continued on next page.

Figure 4.4: *Continued from previous page.*

(A) Relative quantification of iCLIP signal in transcript's exons (5′-UTR, CDS, and 3′-UTRs) (see Methods for details). The target protein is indicated on the right, while the experimental conditions are denoted on top of the plots. Each pointplot shows the median iCLIP signals in exonic regions across three groups of transcripts that are differentially regulated by LIN28A phosphorylation—DOWN, Control, and UP; the error bars represent the interquartile ranges. (B) Line plots show the relative distribution of iCLIP signal (Methods) across 5′-UTRs (left), CDS (middle), and 3′-UTRs (right) for iCLIPs of PABPC1 and PABPC4 in the FGF2-treated LIN28A knockout cells, with (red line) or without (blue line) the induction of LIN28A expression. The plots on the bottom show the relative distribution of iCLIP signal for PABPC1 in naïve cell state, where native WT LIN28A is expressed (green). On each plot, the distributions are visualised for two groups of transcripts: those downregulated upon LIN28A phosphorylation (DOWN, solid line) and those upregulated or unchanged upon LIN28A phosphorylation (UP + Control). (C) Line plot indicates the mean expression-normalised crosslink coverage in the region 200 nts upstream and 50 nts downstream of PAS for iCLIPs of PABPC1 (top) and PABPC4 (bottom) in LIN28A knockout cells with (red line) or without (blue line) induction of LIN28A-WT in the presence of FGF2. Included 3′-UTRs were filtered for minimum length and expression level as described in Methods. Shaded areas indicate a 95% confidence interval. (D) Boxplot shows the count of AAA 3-mer in a region 100nts upstream of PAS. Included genes were filtered as described in Methods to yield 1002 genes in DOWN, 2126 genes in Control, and 778 genes in UP group. Two-sided Mann-Whitney-Wilcoxon test was performed to assess the significance of difference in groups' means; the comparisons and p-values are annotated on the graph.

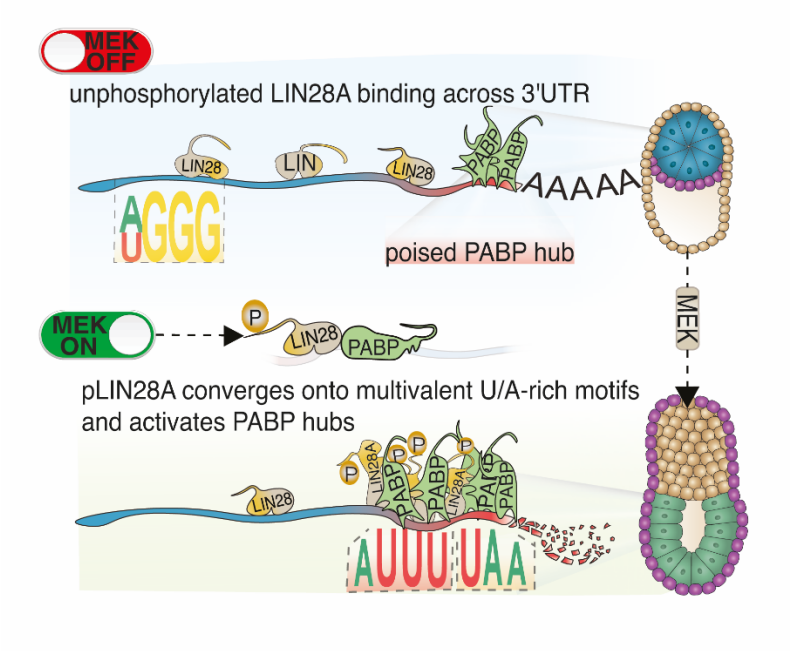


Figure 4.5: Proposed mechanism of LIN28A-mediated mRNA destabilisation in naive-to-primed transition.

4.4 Summary

In this chapter, we explored how post-translational IDR modification in LIN28A drives naïve-to-primed cell fate transition in early embryonic development by promoting destabilisation of naïve mRNA regulon. Below is a summary of how our analyses addressed the hypotheses proposed in the introduction.

Hypothesis 1: Phosphorylation of LIN28A in its IDR changes its RNA binding properties to activate its function in selective mRNA decay. Our results show that the phosphorylation of LIN28A in its C-terminal IDR promotes its 3′-UTR interactions and increases its binding to A/U-rich regions, rather than canonical CSD-bound GAU motifs and zinc-finger-bound WGG motifs. Further, we show that after phosphorylation, LIN28A assembles specifically at the 3′-UTR start and end regions, as compared to dispersed binding along the 3′-UTR seen prior to phosphorylation. We observed an imbalance in pLIN28A localisation on 3′-UTRs with respect to transcript stability: DOWN transcript had a higher proportion of pLIN28A localised to the 3′-UTR termini, compared to 3′-UTR starts, than UP transcripts. When we quantified LIN28A binding to the 3′-UTR terminal regions, we observed that pLIN28A accumulation occurs approximately 40nts upstream of PAS, with a higher level of pLIN28A binding observed in DOWN transcripts. Together this confirms our hypothesis that the IDR phosphorylation of LIN28A changes its RNA binding properties to activate its function in selective mRNA decay.

Hypothesis 2: The transcripts that undergo selective decay exhibit distinctive sequence features that are selectively recognised by phosphorylated LIN28A. Our study reveals that DOWN transcripts exhibit multivalent A-rich and U-rich sequences in the 100nt region before PAS at 3′-UTR termini. This increased multivalency corelates with increased binding of LIN28A to these regions and transcript destabilisation, thus supporting our hypothesis. To establish direct evidence that the increase in A-rich or U-rich motifs in this region promotes mRNA destabilisation in naïve-to-primed transition, further experiments

are needed where the effect of varying levels of these motifs on RNA stability is assessed in form of a screen.

Hypothesis 3: We hypothesise that phosphorylated LIN28A acts on the transcripts in concert with cytoplasmic PABPs to regulate selective mRNA decay in naïve-to-primed transition. Our research supports this hypothesis, by showing that PABPC binds to the 3′-UTR regions to which LIN28A relocates after phosphorylation. We show that DOWN transcripts exhibit more PABPC binding at 3′-UTR termini, which is further enhanced by LIN28A expression. This suggests a cooperative relationship between pLIN28A and cytoplasmic PABPs in mediating transcript stability, thus enabling precise coordination and timing of cell differentiation in naïve-to-primed transition.

4.5 Contributions

All experimental data analysed in this work—including the 3′-end sequencing experiments and iCLIP experiments—were performed by my colleague and supervisor Miha Modic, who also provided guidance and counsel on data analysis. All the bioinformatic analyses presented here were performed by me.

Chapter 5

Discussion

Understanding which motifs RBPs recognise and bind to *in vivo* can elucidate mechanisms through which they regulate gene expression and downstream cellular processes. Despite motif analysis being crucial in characterising protein-RNA interactions, the efforts in describing binding motifs have focused primarily on *in vitro* data (Dominguez et al., 2018; Ray et al., 2009). Motifs obtained from *in vitro* assays represent the nucleotide sequence or RNA structures with the highest affinity to the recombinant RBP in highly controlled conditions: *in vitro,* there is no competition or cooperation with other co-factors, the crowding of biological macromolecules is low, and randomisation and limited length of RNA sequences avoids the assembly on long multivalent RNA regions. While *in vitro* insights importantly contribute to our understanding of RBP binding in cells, they alone cannot explain the binding of RBP *in vivo*.

This discrepancy was highlighted by a recent study, which evaluated the impact of genetic variation in genomic regions corresponding to *in vitro* motif, CLIP peak, or CLIP peak with *in vitro* motif in the vicinity (Romo et al., 2023). They found that genetic variations in CLIP peaks associated with the *in vitro* motif had the highest impact on gene expression and phenotype, and that these regions were under the highest selective pressure; conversely, genetic variation in regions corresponding to the *in vitro* motif in the absence of *in vivo* binding information had the lowest effect on gene expression and phenotype, as well as the lowest selective pressure. These findings indicate that while *in vitro* motifs are valuable in assessing the relevance of *in vivo* binding sites, they alone are not very informative of biologically relevant RBP binding sites, as they capture many sites that the RBP of interest does not bind. Moreover, *in vitro* motifs fail to capture RBP binding that arises from dynamic interactions promoted by the modular structure of RNA-binding domains coupled with flexible IDRs, which were shown to facilitate RBP binding on long multivalent RNA regions through RBP oligomerisation (Hallegger et al., 2021).

In this work, we develop PEKA, a computational tool that enables accurate and nuanced motif analysis of RBP binding sites *in vivo*, by providing the options to analyse motifs in specific transcriptomic regions and within the context of repetitive elements. The outputs of PEKA are easily applied to comparative studies of RBP binding specificities. Utilising these comparative studies in different contexts, we were able to gain a global understanding of the biological and technical factors that impact the detection of RBP motifs from CLIP data (Kuret et al., 2022). Furthermore, we delved deeper into the modulation of a single RBP, LIN28A, through post-translational modification to understand how such regulation of RBP influences the process of selective mRNA decay, a crucial mechanism for promoting pluripotency progression in early embryonic development (Modic et al., 2023). In essence, this work presents a comprehensive approach

to the investigation of RBP specificity, integrating technical, biological, and functional perspectives.

5.1 Contributions of the Study

Briefly, our work explores the intricate molecular grammar underlying protein-RNA interactions through the lens of sequence motifs, with the focus on the crosstalk between canonical RNA-binding domains and IDRs. It presents a computational tool for motif analysis from CLIP data, established comparative analyses of enriched motifs from CLIP data as a valuable approach to gain insights into data quality and into the biology of RBP specificity and function.

1. *Development of an accessible computational tool for motif discovery from CLIP data.* We developed positionally-enriched k-mer analysis, i.e. PEKA—a computational approach for motif-discovery from CLIP data, which leverages precise positional information of crosslink sites to find enriched motifs, while simultaneously minimises the impacts of sequence biases at crosslink sites or in genomic regions on enriched motifs. Unlike other methods, PEKA uses low-count crosslinks within the same CLIP experiment as background, making it useful for a range of CLIP experiments without requiring external controls. We show that PEKA recovers relevant motifs in eCLIP, iCLIP and PAR-CLIP experiments, and does so at the same level as the related mCross method, when benchmarked on eCLIP (Figure 2.4A). However, PEKA presents several improvements over mCross. For example, while mCross generates PWMs, the results of PEKA are presented as k-mer clusters, which enables it to visualise the more complex and mechanistically relevant enrichment patterns (see Figure 2.2 for PEKA motifs, and Figure 2.3 for mCross motifs), and is also more compatible with further cross-RBP comparative visualisations of motifs. Furthermore, PEKA enables regional analysis and visualisation of enriched motifs, allowing the user to distinguish between motifs specific to a particular genomic region for the investigated RBP. Most importantly, PEKA code is available open source via GitHub and Bioconda in an accessible and user-friendly manner with adjustable parameters. It is also integrated into the Flow platform and the *nf-core/clipseq* pipeline for reproducible analysis of CLIP data (Capitanchik et al., 2023; West et al., 2021).
2. *Variants of CLIP methods have distinct technical biases, which are reflected in enriched motifs.* By comparing global trends in enriched motifs for distinct RBPs between CLIP data and *in vitro* data, we discovered that eCLIP and PAR-CLIP method have distinct sequence biases associated with crosslink sites (Figure 3.3C, Figure 3.5D). These biases likely reflect a combination of crosslink bias and regional sequence biases. We show that in both eCLIP and PAR-CLIP, PEKA reduces these biases. Our comparison of PEKA with mCross on eCLIP data revealed how different motif discovery tools, that use different strategies to model background and foreground in CLIP experiments, have distinct effects on enriched motifs. While PEKA has a slight positive bias towards U- and G-rich motifs, mCross has a strong negative bias against U and AU-rich motifs (Figure 3.3C). This is consistent with our finding that PEKA outperforms mCross for RBPs that bind to U-rich motifs, and conversely, mCross outperforms PEKA for RBPs that do not bind strongly to U-rich motifs (Figure 2.4C).
3. *Over a third of eCLIP data show low specificity of enriched motifs.* At the time of this study, eCLIPs collected on ENCODE represent the largest collection of CLIP

experiments and was leveraged to study *in vivo* binding preferences of RBPs (Feng et al., 2019; Katsantoni et al., 2023; Van Nostrand et al., 2020), and train machine learning models (Ghanbari & Ohler, 2020; Horlacher et al., 2023). Despite its importance and widespread application, the systematic assessment of data specificity was lacking. The CLIP experimental protocol omits the visualisation of purified protein-RNA complexes, which is normally used for experimental optimisation of specificity via visual analysis of expected vs. co-purified RBPs (Hafner et al., 2021). Previous studies evaluated successful IP, library complexity, and the number of reproducible CLIP peaks as a signature of data quality (Van Nostrand et al., 2020); however, these metrics serve mainly as measures of data sensitivity (Chakrabarti et al., 2018). As a measure of data specificity, previous studies have reported the agreement of enriched motifs between eCLIP with RBNS (Van Nostrand et al., 2020) but have not used such analyses to compare various quantitative metrics of data specificity across datasets. We find that 5% of datasets contain the highest specificity characteristics of high agreement with orthogonal in vitro data and inter-eCLIP data specificity, whereas ~33% of datasets have low inter-eCLIP specificity, lack orthogonal in vitro data, and in most cases show a predominance of G-rich motifs (Figure 3.9A). Notably, such motifs are found enriched even in some eCLIP (but not iCLIP or PAR-CLIP) datasets from the highest quality datasets, such as TIA1 (Figure 3.1). It has been hypothesised that these motifs may be bound by co-purified RBPs (Van Nostrand et al., 2020), and our analyses indicate that this contamination affects mainly eCLIP, but not iCLIP and PAR-CLIP datasets of the same proteins (Figure 3.2). Since we find that enrichment of G-rich motifs is seen in datasets with preferential intronic crosslinking, we speculate that G-rich background might result from contamination of chromatin and associated RBPs. Such contamination tends to emerge when cell extracts are too concentrated and viscous (Grabski, 2009), and therefore the conditions of CLIP normally recommend sonication and DNase treatment of a well-diluted extract, and visualisation of purified protein-RNA complexes to confirm lack of contaminating signal (Ule et al., 2005). Importantly, RBPs with low sequence specificity and low extent of motif enrichment (low PEKA scores) contain more IDRs and low-complexity regions, as well as in RNA binding domains that are not specialised for recognition of single stranded sequence motifs (Figure 3.9A,C,D). Together these features indicate that these RBPs might lack sequence specificity, but potentially recognise some other features on the RNA; or that they do not bind directly to RNA but are instead associated with it via other protein co-factors through interactions mediated by IDRs and low-complexity regions. A recent study used the RNAcompete assay to probe sequence specificity of ~500 RBPs without canonical RBDs and found that the vast majority of these RBPs indeed lack sequence specificity (Ray et al., 2023).

4. *The use of SMInput controls does not universally improve motif-discovery from eCLIP data.* The use of SMInput controls is considered a key aspect of eCLIP data analysis. As eCLIP protocol skips the gel visualisation step, the RNA from the non-IPed cell lysate isolated from the same membrane region as the IPed protein, i.e. the SMInput control, provides a measurement of non-antigen-specific background that could potentially contaminate the IPed sample (Van Nostrand et al., 2016). Despite its widespread use, we show that SMInput controls do not universally enhance the accuracy of motif finding but do lead to greatly decreased sensitivity (Figure 3.6, Figure 3.7) when used with a stringent arbitrary threshold for enrichment as is employed by narrowPeaks (Van Nostrand et al., 2016). The effectiveness of SMInput filtering depends on the RBP: it can enhance motif and

binding site discovery for RBPs that do not contribute significantly to the signal in the SMInput, but detrimental for highly expressed RBPs that do. Furthermore, antigen-specific background in eCLIP experiment, such as co-purified protein co-factors, would still be enriched over the SMInput control. We argue that the utility of SMInput filtering depends on a variety of factors and thus cannot be viewed as a substitute for controlling for protein and RNA impurities on the gel. Since our study, more nuanced models for peak assignment using SMInput have been developed (Boyle et al., 2023; Schwarzl et al., 2022). These models replace arbitrary cutoffs for enrichment in fixed regions by modelling differential enrichment of crosslinks in eCLIP or SMInput across genomic regions using sliding windows. This allows them to recover more relevant binding sites and significantly improve data sensitivity compared to the narrowPeak approach.

5. *RBPs with canonical RNA binding domains and low number of IDRs and low-complexity regions, show high specificity of enriched motifs and high sensitivity of eCLIP data.* We find that eCLIPs of RBPs with the highest motif enrichments (high mean PEKA scores) and inter-eCLIP data specificity (low similarity score) tend to have more than one canonical RNA binding domain—defined here as RRM or KH domain (Figure 3.9A,B). While this informs on the extent that RBPs might be sequence-specific (Dominguez et al., 2018; Jankowsky & Harris, 2015; Ray et al., 2013), it also informs on the specificity of eCLIP datasets. Datasets that agree best with in vitro data are expected to be most specific (i.e., group 2 in Figure 3.9A), and these have the highest PEKA scores and inter-eCLIP data specificity. The same RBPs are also most efficiently identified by enhanced RNA-interactome capture (eRIC), indicating that they crosslink well. Interestingly, all RBPs in group 2 predominantly bind to introns. Furthermore, RBPs with high extent of motif enrichment and high inter-eCLIP specificity tend to have a lower number of IDRs and low-complexity regions (Figure 3.9B,C,D). Further investigation is required to understand how protein structure elements and regional binding preferences of RBPs affect their RNA interactions and the sensitivity of CLIP experiments. A recent study by Feng et al. found that the UV-C crosslinking (used in eCLIP and iCLIP) is highly specific, requiring certain amino acid residues and types of intermolecular interactions (Feng et al., 2022). This research highlighted the key features enabling crosslinking in canonical RNA-binding domains, such as the KH and RRM. However, the analysis of non-canonical RBPs revealed that their crosslinking is less specific compared to canonical RBPs. These findings align with our study, which observed lower sensitivity in eCLIP data for RBPs without canonical RBDs, suggesting a potential limitation of UV-C crosslinking in capturing the RNA interactions of such proteins. Another study examining the sequence specificity of non-canonical RBPs found that most of them lack sequence specificity (Ray et al., 2023), which corroborates our observation of the prevalence of low-specificity G-rich motifs for RBPs lacking canonical RBDs. These collective findings underscore the limitations of UV-C crosslinking for certain proteins and highlight the need for alternative crosslinking strategies that can capture protein-RNA contacts even when UV-C crosslinking constraints are not met. The use of photo-reactive nucleoside analogues as employed by PAR-CLIP might partially solve this problem, as we do not observe the correlation between eCLIP sensitivity and the content of KH- / RRM- domains in PAR-CLIP (Figure 3.10); but notably the number of thresholded crosslinks in PAR-CLIP is generally lower than in eCLIP (Figure 3.9A, Figure 3.10). Taken together, additional *in vivo* studies of non-canonical RBPs are necessary, together with stringent controls, to determine the nature of their interactions with RNA—whether they are direct or mediated

through other protein effectors. Moreover, the effectiveness of different crosslinking approaches for these RBPs must be evaluated.

6. *Phosphorylation of LIN28A by MEK/ERK signalling cascade changes its RNA interactions, mediating the selective decay naive mRNAs to promote cell differentiation in early embryonic development.*

   The phosphorylation of LIN28A by the MEK/ERK signalling cascade is a key event in early embryonic development, which promotes cell transition from naïve to primed pluripotency by inducing selective clearance of naïve regulon mRNAs. This post-translational modification alters the RNA interactions of LIN28A, enhancing its interactions with 3′-UTRs (Figure 4.1E), particularly at the 3′-UTR termini approximately 50 nucleotides upstream of PAS (Figure 4.3A,B). Interestingly, these sites are occupied by PABPC even in the absence of LIN28A expression and prior to MEK/ERK activation (Figure 4.4B,C). Our study reveals that the selectivity of pLIN28A-mediated mRNA decay is conferred by the following features: downregulated mRNAs have a higher density of AU-rich motifs upstream of PAS (Figure 4.3C,D), and these regions also exhibit a higher density of PABPC binding (Figure 4.4C). Upon LIN28A expression and phosphorylation, the number of PABPC1 and PABPC4 bound to these regions further increases (Figure 4.4C). It is likely that RNA sequence with high valency of polyA motifs enables greater initial accumulation of cytoplasmic PABPs at these regions (Figure 4.4D), which might in turn recruit more LIN28A to these sites after MEK/ERK activation (Figure 4.3B), leading to their degradation. These findings fit into the broader context of how interactions between RNA, RBPs and effector proteins tightly control RNA metabolism, enabling the same effectors to exert diverse actions (He et al., 2023; K. Shah et al., 2023). For example, cytoplasmic PABPs can either stabilise or destabilise mRNAs and promote or repress translation, depending on its interaction network (Qi et al., 2022). A recent study showed that effector proteins PABPC1, PABPC4, and CCR4-NOT are controlled by the RBP Unkempt to mediate translational repression of mRNAs. Unkempt interacts with cytoplasmic PABPs and CCR4-NOT via short amino acid motifs in its IDR, thus bringing together the relevant RNA targets which contain sequence motifs recognised by Unkempt, i.e., substrates, and the effector proteins that act on them (K. Shah et al., 2023). A similar mechanism of CCR4-NOT recruitment could be occurring in the case of LIN28A-mediated decay, but instead of mediating only translational repression, pLIN28A might activate the deadenylase activity of CCR4-NOT, leading to mRNA decay. Contrary to Unkempt, LIN28A's IDR does not contain PAM2 motifs that mediate interaction with cytoplasmic PABPs (R. Yang et al., 2011), pointing to a different mechanism of cooperation. Further experiments are required to understand whether the interactions between cytoplasmic PABPs and LIN28A are direct or indirect, and to identify other protein cofactors that play a role in pLIN28A-mediated mRNA decay. Nevertheless, our study provides valuable insights into how regulator and effector RBPs function together on the RNA to control its metabolism, and how this interaction network can be rapidly modulated through post-translational modification of IDRs. In summary, our study provides valuable insights into the role of LIN28A phosphorylation driving pluripotency progression and explains features that mediate selective decay of naïve pluripotency mRNAs. These insights could potentially expand our understanding of the pathological roles of LIN28A and lead to identification of novel therapeutic targets and approaches.

7. *K-mer analysis coupled with clustering provides a robust framework to study changes in RBP-RNA interactions across diverse cellular conditions, and to detect*

common motif enrichment patterns that emerge in different variants of CLIP method and inform on technical biases. Our study utilised a comparative analysis of enriched k-mers from PEKA across a range of RBPs, CLIP methods, and cellular conditions. This approach yielded technical insights into the impact of different CLIP methods on motif enrichment, as well as biological insights into how a post-translational modification of LIN28A triggers a shift in its specificity and activates its role in mRNA decay. In the case of LIN28A, we demonstrated that k-mer enrichment profiles robustly cluster together replicates from identical conditions. We anticipate that similar analyses across a variety of cell lines and conditions could reveal additional shifts in specificity driven by RBP modifications.

5.2 Limitations of the Study

Despite the valuable contributions of our study, it is important to acknowledge the following limitations:

1. *In contrast to eCLIP data, PEKA discovered fewer relevant motifs from PAR-CLIP data, when compared to in vitro derived motifs (Figure 3.5B).* Factors that potentially hinder PEKA's performance on motif discovery in PAR-CLIP datasets might relate to lower sensitivity (as indicated by the number of tXn), but higher specificity of the data (Figure 3.10), due to the stringency of IP. In both eCLIP and PAR-CLIP, we observe a correlation between PEKA-score and the number of thresholded crosslinks; in eCLIP PEKA score also correlates strongly with recall, whereas in PAR-CLIP, this correlation is not evident. This suggests that the crosslinking approach used by PAR-CLIP decreases the relationship between cDNA counts of crosslinking events, and greater reliability of motif detection. In PAR-CLIP, only the photoreactive 4SU can crosslink, and this photoreactive analogue competes for its integration into the RNA with the native uridine. Therefore, the number of transitions at a certain genomic position does not only relate to the RBP binding at that position, but also to the frequency of 4SU integration at this position. As we did not benchmark the performance of other motif discovery tools on PAR-CLIP data, we cannot disentangle whether the poorer agreement of motifs with *in vitro* data is due to PEKA being a sub-optimal approach to motif discovery, or whether motifs discovered by PEKA indeed occur at binding sites identified by PAR-CLIP and the data itself lead to lower agreement with motifs derived from *in vitro* data. Nevertheless, for various PAR-CLIP datasets analysed here, the motifs have been reported in the literature and can therefore be compared to PEKA enrichments. For example, in Figure 3.2 PEKA the recall for PAR-CLIP of HNRNPC is low, however the purine-rich motifs recovered by PEKA are similar to the GAAC motif, which was reported by the original study as enriched in m6A modified reads bound by the HNRNPC (Liu et al., 2015). In the case of TAF15, EWSR and FUS, the original study did not find any enriched motifs for these proteins (Hoell et al., 2011), however when analysed with PEKA, the top motifs for FUS and TAF15 do show partial agreement with *in vitro* data. For IGF2BP1/2 and PUM2 PARCLIP had high agreement with *in vitro* data, compared to eCLIP, and the motifs are in agreement with those reported by the original study, which used a different motif discovery method: CAU motifs for IGF2BP1/2 and UGUA-AUA motifs for PUM2 (Hafner et al., 2010). These comparisons highlight that the motif discovery performance in PEKA reflects the heterogeneity of PAR-CLIP datasets. To accurately evaluate the performance of motif discovery tools on PAR-CLIP data, and sequence biases that affect enriched motifs in PAR-CLIP, a large

compendium of PAR-CLIP experiments should be produced for diverse RBPs, while minimising technical variation.

2. *Our study does not consider RNA structure in motif discovery.* A recent study analysed binding specificity of 144 distinct RBPs, focusing on their preference for either RNA sequence and structure (Laverty et al., 2022). The study found that 90% of the analysed RBPs preferred unstructured RNA regions, which allow access to linear sequence motifs (Laverty et al., 2022). However, it is important to note that the RBPs in this study mainly contained canonical RNA binding domains, and the specificity of non-canonical RBPs may differ.
   Another study aimed to characterise the binding preferences of non-canonical RBPs by evaluating sequence motifs and structural preferences for a subset of eCLIP data targeting these RBPs (Ray et al., 2023). The authors reported that they could not derive any significant motifs for 12 out of 31 eCLIPs. In contrast, 17 eCLIPs did yield motifs, but the motifs were very similar to each other and showed little or no preference for RNA structure. However, this lack of sequence or structural specificity may be due to the lower quality of eCLIP data, as previously discussed. Therefore, the obtained motifs may not accurately reflect the true binding determinants of the tested RBPs.
   Despite these limitations, these studies suggest that analysing RBP binding specificity in terms of linear sequence motifs—as done in our study—is likely relevant for most RBPs, especially those with canonical binding domains. However, for the few RBPs that primarily recognise RNA structure, PEKA may not be able to identify relevant motifs. Therefore, future studies could benefit from using PEKA in combination with other motif discovery tools that can inform on structural motifs of RBPs, particularly as more unconventional RBPs are studied. Finally, we caution that even tools that model structure may not accurately identify the structural motifs, as they are limited to predicting local mRNA structures that may not occur in the cellular milieu and cannot consider distal or inter-molecular RNA structures that may occur within cells.
3. *For comparative analysis of eCLIP and PAR-CLIP data across diverse RBPs we only considered RNA regions found in protein-coding genes.* Our analysis of motif enrichments is focused on protein-coding genes because most eCLIP datasets correspond to RBPs that do not bind to non-coding RNAs (ncRNAs), and partly because the RBPs that do, primarily bind to a few abundant ncRNAs (Van Nostrand et al., 2020) that are not sufficient for motif derivation. Moreover, abundant ncRNAs are highly structured and modified and involve complex and multi-step RNP assembly mechanisms, and therefore require integrative analysis of multiple types of data and structural modelling before the functional sequence motifs can be extracted (Maticzka et al., 2014; Zampetaki et al., 2018). Nevertheless, a subset of eCLIP data is highly enriched in ncRNAs, and clustering of eCLIP by regional crosslinking profiles that include abundant ncRNAs and repetitive RNA elements does lead to smaller clusters (Van Nostrand et al., 2016) compared to the clusters defined by enriched motifs in our study (Figure 3.8). Therefore, it is important to note that the similarity index from our study only represents specificity of motifs enriched in protein-coding genes, rather than specificity of eCLIP data as a whole. In the future, it will be valuable to analyse enriched sequence and structural motifs in combination with the types of bound RNA to understand additional features that contribute to the specificity of CLIP datasets.
4. *We do not prove causality between characteristic features of DOWN transcripts and their decay.* Our study elucidates the features that correlate with pLIN28A-

mediated mRNA decay; however, it does not prove causality. To understand whether AU-rich sequences are indeed essential to activate LIN28A-mediated decay, it is important to follow up with orthogonal studies which perturb these motifs, and measure how perturbations affect transcript stability and RBP binding in that region.

5. *Our study primarily focused on a subset of transcripts that were downregulated following LIN28A phosphorylation.* It is important to note that upon LIN28A phosphorylation many transcripts were upregulated. This upregulation could result from phosphorylated LIN28A promoting the transcription of these genes or enhancing the stability of mRNAs. Interestingly, we observed that upregulated mRNAs exhibited distinct features compared to the downregulated transcripts. For instance, they had a lower content of AU-rich motifs upstream of PAS (Figure 4.3D) and showed reduced binding of cytoplasmic PABPs to these terminal 3′-UTR regions in the LIN28A KO condition (Figure 4.4C). Additionally, we found that upregulated transcripts had a greater accumulation of PABPC binding around the STOP codon, suggesting a potential role in translation (Figure 4.4B). While a comprehensive analysis of upregulated transcripts was beyond the scope of this work, it would be interesting to consider this aspect of gene expression regulation in future studies.

5.3 Future Directions

We anticipate that recent advancements in high-throughput methodologies for profiling protein-RNA interactions will generate an unprecedented volume of CLIP data for various proteins (Wolin et al., 2023). This development provides an opportunity for detailed mechanistic studies of complex biological processes. Our work demonstrated how comparative analyses of CLIP data can be leveraged to derive novel biological insights and assess the specificity and sensitivity of CLIP data. Emerging techniques that enable simultaneous probing of multiple RBPs could significantly enhance our understanding of intricate regulatory networks, where multiple protein and RNA molecules work in concert (He et al., 2023).

However, this expansion also underscores the need for rigorous quality control and a thorough understanding of potential methodological biases. For methodologies like SPIDR, where gel visualisation is not possible, and IP conditions and RNA digestion cannot be optimised for each protein (Wolin et al., 2023), quality control through motif analysis becomes particularly useful. A potential strategy could involve IPing RBPs with known motif preferences and analysing these in conjunction with other IPed proteins to identify non-specific motif signatures and flag unsuccessful experiments.

To facilitate the comparative analysis of CLIP data in the future, PEKA has been integrated with the flow.bio web platform (Capitanchik et al., 2023), a tool designed for the analysis of high-throughput biological datasets. As the volume of datasets continues to grow, platforms that enable reproducible analysis, identification, storage, and machine readability will become increasingly important. This aligns with the FAIR (Findable, Accessible, Interoperable, and Reusable) data principles, which aim to enhance the ability of machines to automatically find and use the data, as well as support its reuse by individuals (Wilkinson et al., 2016). In terms of CLIP data, we hope that flow.bio will emerge as a leading platform, fostering the creation of a large, public database resource. This would not only streamline data analysis and enable robust quality control, but also promote collaboration and knowledge sharing within the scientific community.

Besides enhancing data analysis and constructing a comprehensive database of CLIP datasets, it is equally important to advance the technical aspects of the methodologies used. A recent study has elucidated the limitations of UV-C crosslinking, noting that stringent structural constraints in protein-RNA contact must be met for crosslinking to occur (Feng et al., 2022). While this results in high-specificity crosslink sites for canonical RBPs that form such interactions with the RNA, unconventional RBPs, particularly those with highly disordered structures, may not crosslink as effectively using the UV-C crosslinking approach. Currently, UV-C crosslinking is the standard for most CLIP methods, apart from PAR-CLIP, which employs crosslinking with photoactivatable ribonucleoside analogues (F. C. Y. Lee & Ule, 2018). As more unconventional RBPs are studied, there is a growing need to design more inclusive crosslinking approaches. These should be capable of capturing zero-length protein-RNA interactions with high sensitivity and without significant off-target effects. At present, crosslinking strategy utilising photoreactive nucleoside analogues might be able to address some of the limitations of UV-C crosslinking. However, due to a lack of diverse datasets with low batch variation, it is unclear whether the benefits of PAR-CLIP outweigh its drawbacks, such as the potential off-target effects of photoactivatable ribonucleoside analogues (Altieri & Hertel, 2021; Burger et al., 2013; Eshleman et al., 2020) and lower sensitivity, as shown in this work. We hope future studies will tackle these issues, and lead to the development of alternative crosslinking strategies that would be sensitive and effective for a broad scope of RBPs.

In addition to insights into the biases of CLIP methods and sequence specificity of RBPs globally, our work shed light on the complex interplay between LIN28A, cytoplasmic PABPs and RNA that promotes naïve-to-primed cell-fate transition in early embryonic development, by mediating selective decay of transcripts that encode naïve pluripotency factors. We showed that the protein-RNA regulatory networks driving this process are modulated by phosphorylation of LIN28A in its IDR. Given the involvement of LIN28A in diverse pathologies (Wu et al., 2022)—and particularly cancers (Balachandran & Narendran, 2023; H. Wang et al., 2016)—understanding its regulatory role in RNA stability could provide valuable insights into these diseases. Future studies should therefore explore the full network of regulator and effector proteins required for pLIN28A-mediated decay and how they act on the mRNA to promote decay. It would also be intriguing to investigate whether the poised PABPC hubs at 3′-UTRs, that can be modulated by different RBPs to achieve distinct effects on mRNA stability, are a widespread phenomenon. Further research in this area could provide a useful framework for understanding the forces governing mRNA stability more generally, and could help explain the effects of LIN28A and PABPC dysregulation in pathologies.

Finally, our research underscores the potential roles of IDRs in RBPs as vital modulators of cellular processes, that enable rapid response to stimuli by engaging diverse RNA-related pathways, such as RNA stability, translation, transcription, RNA processing, etc. This raises broader questions about the role of IDR modifications and perturbations in fine-tuning cellular processes through RBPs. To understand their functional significance, it is crucial to investigate post-translational modifications and other perturbations of IDRs *in vivo*. Our work serves as a foundation for understanding these complex interactions through bioinformatic analysis of CLIP data. As more dynamic biological data becomes available and with development of crosslinking technologies, future studies should explore these principles to expand our understanding of the functional roles of IDRs in RBPs.

Chapter 6

Methods

6.1 General

This section describes the general methods that were applied more than once over the course of this work. The descriptions are adapted from our publication (Kuret *et al.*, 2022).

6.1.1 Peak calling with Clippy

For motif analyses in this work, peaks of crosslinking events were determined using Clippy with default parameters and *--intergenic_peak_threshold* set to 5 to enable intergenic peak calling (Capitanchik et al., 2022). Clippy is optimised for fast CLIP peak calling informed by genomic annotations. First, the single-nucleotide crosslink signal is smoothed using a rolling mean of window size 10. Next, to determine the minimum height thresholds for peaks, the genome is split into regions based on annotations and crosslink density. For our analysis, genes with more than five mapped cDNAs were split into exons and introns; intergenic regions were defined based on smoothed crosslink density. Clippy determines peak summits with the *find_peaks* function from the *scipy* Python library (Virtanen *et al.*, 2020) and filters them based on their height removing the summits below the mean of cDNA signal in the region. Additionally, summits with a prominence below this regional mean are discarded. Broad peak regions are calculated for the remaining summits based on the shape of the surrounding signal curve; finally, these peak regions are filtered to contain a minimum of five cDNAs.

For the analyses presented in Chapters 2 and 3, we used v1.4.1; for analyses presented in Chapter 4, we used v1.5.0.

6.1.2 PEKA

PEKA is a tool for finding enriched sequence motifs from CLIP data available at GitHub (https://github.com/ulelab/peka) and Bioconda (https://anaconda.org/bioconda/peka); the code used in this work is also archived on Zenodo (A. G. Amalietti et al., 2022). The first step of PEKA is thresholding, which splits the crosslinks into high confidence thresholded crosslinks (tXn) and reference background crosslinks (oXn). To obtain tXn, each transcript is considered separately, such that all exons (including CDS, UTRs, and ncRNAs) within a gene are combined into one region, and each intron and intergenic region are treated as their own region. Within each region, a cDNA count threshold is determined at which ≥70% (the default, but this percentile can be modified by the user) of the crosslink sites within the region have a cDNA count equal or below the threshold; e.g., if the region contains 10 crosslinks, nine out of which have a cDNA count of 1 and one has a cDNA count of 2, the threshold for that particular region is set to 1. tXn are then identified that

have cDNA count above the threshold and overlap with the peaks that are provided by the user. oXn are defined as all crosslinks that fall outside of peaks. The impact of cDNA count threshold will vary based on the depth of the analysed CLIP library. Currently available CLIP datasets are dominated by crosslinks identified by a single cDNA: setting the cDNA count threshold to 70% generally serves to remove the crosslinks with a score of 1 from tXn, as such crosslinks represent most of the data. Increasing the threshold above 70% decreases the number of identified tXn, which can be detrimental especially for small samples. Conversely, if the cDNA count threshold is set to 0, then all crosslinks within peaks will be used as tXn, which can be beneficial for very small samples, but otherwise generally leads to decreased specificity of enriched motifs by ignoring the quantitative information available from cDNA counts. With increasing sequencing depth, we expect that this threshold will become more important in distinguishing between crosslinks with higher cDNA counts representing various levels of binding strength. For data with abundance of crosslinks, we randomly sample tXn and oXn to obtain 1 million tXn and 3 million oXn positions; this decreases memory requirements and time of computation, while yielding results that are comparable to using all tXn and oXn positions. While this sampling is done by default, it can be turned off by the user.

By default, PEKA examines motif enrichment in the following transcriptomic regions: introns (from both coding and non-coding genes); 3′-UTRs; other protein-coding exon regions (comprised of coding sequence exons and 5′-UTR), non-coding RNAs (comprised of exons from non-coding RNAs); intergenic regions; protein-coding genes (comprised of full sequence of protein-coding genes); and whole genome. The segmentation of genomic regions is a required input for PEKA analysis; it can be produced from either GENCODE or Ensembl annotation with the *get_segments* function in the iCount tool (Curk, 2019). For this study we prepared the regions file from the GENCODE primary assembly V39 annotation, which we filtered before segmentation to ensure robust assignment of genomic regions; we removed transcripts with low support level (less than 2) in genes where such transcripts were available.

In addition to motif discovery in various genomic regions, PEKA supports the use of repeat-masked genomes providing the following options: (1) excluding repeat elements from motif discovery (used in this study); (2) conducting motif discovery separately in repeat elements and non-repeats; (3) conducting motif enrichment exclusively in repeat elements (4) not discriminating between repeat elements and non-repeats (the default).

Foreground and background sequences for motif discovery are extracted around tXn and oXn, respectively, from the user-provided genome file. Both foreground and background sequences are extracted within the transcriptomic region that is currently analysed, to minimise the impact of regional sequence biases on enriched motifs. Foreground sequences span −150…150nt around tXn and are subdivided into the proximal (−20…20nt around tXn) and distal window (−150…−100 and 100…150nt around tXn); motif discovery is conducted within a proximal window, while distal windows are needed to define the threshold for relevant positions, as is described in the following paragraph. Background sequences consist only of the proximal window and span −20...20nt around oXn. The values used for proximal and distal windows used in our study are set by default in the code but can be adjusted by the user. However, we recommend that the selected proximal window is not more than −50...50nt, as we rarely see enrichment of relevant motifs further than 50nt from crosslink sites.

For each k-mer, PEKA scans across the foreground and background sequences and records the presence of a k-mer by assigning a count of one if present or zero if absent (Figure 2.1D); for k-mers of odd lengths, the position that coincides with the middle of the k-mer is assigned the count, and for k-mers of even lengths, the position that corresponds to its length divided by a factor of 2 is assigned the count (i.e., 3rd overlapping nucleotide

for a 6mer). For sequences located in each transcriptomic region, mean k-mer counts are calculated at each position to obtain k-mer occurrences around crosslink sites. Afterwards, relative k-mer occurrence is calculated by dividing k-mer occurrence at each position with the mean k-mer occurrence across all positions within distal windows of the foreground sequences. Relative k-mer occurrence is calculated separately for the foreground sequences (to get RtXn) and for 100 samples of randomly selected background sequences, for which the sample size corresponds to the number of foreground sequences (to get 100 RoXn distributions).

PEKA calculates the enrichment of each k-mer at sequence positions where the k-mer is present above a background level, i.e., "the relevant sequence positions" (Figure 2.1E). These positions are identified by analysis of the relative occurrences, such that each sequence position within the proximal window gets its own threshold value, calculated from the 100 RoXn distributions that represent the background. For each position, the calculated RoXn values of all possible k-mers (100×4^k values) are combined into a union and then the threshold is defined as the value at which a specified percentile of evaluated RoXn is lower than that value; by default, this percentile is set automatically based on k-mer length and the number of foreground sequences, but a specific value can also be passed by the user. For each k-mer, the positions at which RtXn exceeds the threshold are marked as relevant. Setting a higher percentile will result in a higher threshold and fewer relevant positions, while setting the percentile to zero will result in all positions within a proximal window being considered as relevant.

Finally, the PEKA-score is calculated by evaluating k-mer frequency at the relevant sequence positions in the foreground compared to the background (Figure 2.1E,F). For each k-mer, a mean RtXn (ARtXn, i.e., estimated occurrence around tXn) is calculated across relevant positions to evaluate its foreground frequency. To evaluate k-mers' background frequencies, mean RoXn (ARoXn, i.e., estimated occurrences around oXn) are calculated across relevant positions for 100 random samples of oXn. Finally, PEKA-score is calculated as:

$$PEKAscore = (ARtXn - mean(ARoXn))/SD(ARoXn) \quad (6.1)$$

where $mean(ARoXn)$ represents the mean of ARoXn values across all 100 samples, and $SD(ARoXn)$ represents the standard deviation of ARoXn values across all 100 samples (Figure 2.1F).

The k-mers are ranked by PEKA-score from the most to the least enriched, and all results are given in a table. In addition, a p-value is calculated for each k-mer from a distribution of PEKA-scores across the dataset. For this, PEKA-scores are first standardised by subtracting the mean and dividing by the standard deviation across all scores; then, p-values are obtained for each k-mer with the right-sided tail test. PEKA-scores for all datasets analysed in this study are available in supplementary materials of our paper ((Kuret *et al.*, 2022), Additional file 5: Table S4).

To represent enriched motifs, PEKA visualises occurrence profiles of the top n k-mers clustered by sequence characteristics and occurrence distributions (Figure 2.1G); by default, the n is set to 20, but this can be modified by the user. Sequence similarity between k-mers is represented as a matrix of pairwise Jaccard indices calculated with the Python *textdistance* library (*TextDistance*, 2021). The similarity of k-mer occurrence distributions is measured with (1) Spearman rank correlation coefficient, (2) occurrence maximal values, metrics are combined at varying weights, and the resulting matrix is used for clustering; the optimal clusters are selected for visualisation based on the lowest standard deviation of occurrence medians within clusters.

For the analyses presented in Chapter 3, we used PEKA v0.1.6—deposited to Zenodo (A. G. Amalietti et al., 2022)—and focused on the analysis of crosslink sites in the protein-coding regions, consisting of CDS, UTRs and introns. For motif discovery, we merged crosslink sites from replicate CLIP experiments by summing up cDNA counts at respective positions; the merged experiments are indicated in supplementary materials of our paper (Kuret et al., 2022); see Additional file 3. For these analyses, we used the GRCh38.p12 primary assembly genome, with all annotated repeat sequences soft-masked using RepeatMasker v4.1.0. For analyses presented in Chapter 4, we used PEKA v1.0.0 and analysed each replicate CLIP experiment separately. The full list of PEKA settings that were applied for analyses presented in Chapter 3 and Chapter 4 are summarised in Table 6.1 and Table 6.2, respectively.

Table 6.1: PEKA settings applied to analyses presented in Chapter 3.

Script flag	**Parameter description**	**Setting**
-i	path to CLIP peaks in BED6 file format	path to CLIP peaks in BED6 file format, determined with Clippy
-x	path to CLIP crosslinks in BED6 file format	path to CLIP crosslinks in BED6 file format
-sr	genomic regions in which to perform motif enrichment	whole_gene, intron, other_exon, UTR3
-k	k-mer length	5
-g	genome fasta file	a soft-masked fasta file for GRCh38.p12, obtained from GENCODE
-gi	genome fasta index file	a fasta index file for GRCh38.p12, obtained with "faidx" function from samtools
-r	genome segmentation file in GTF format	produced with iCount segment function from pre-filtered GRCh38 GENCODE v39 primary assembly anotation
-s	length of the smoothing window	6
-p	percentile threshold for determination of thresholded crosslinks	0.7
-w	window around thresholded crosslinks for finding enriched kmers	20
-dw	distal window around enriched kmers to calculate relative k-mer occurrence	150
-n	number of top-ranked k-mers to cluster and visualise	20
-c	maximum number of k-mer clusters	5
-re	How to treat repeating regions within genome.	remove_repeats
-pos	percentile to set as threshold for relevant positions	None
-relax	Whether to relax automatically calculated threshold for relevant positions.	True

-sub	Whether to subsample crosslinks for motif enrichment, if there are many crosslinks in the dataset.	True
-a	Whether to save all outputs.	True
-seed	Set the seeds for random sampling to ensure reproducibility. For each random sample (n=100), the seed incrementally increases from 0 to 99.	True

Table 6.2: PEKA settings applied to analyses presented in Chapter 4.

Script flag	Parameter description	Setting
-i	path to CLIP peaks in BED6 file format	path to CLIP peaks in BED6 file format, determined with Clippy
-x	path to CLIP crosslinks in BED6 file format	path to CLIP crosslinks in BED6 file format
-sr	genomic regions in which to perform motif enrichment	UTR3
-k	k-mer length	5
-g	genome fasta file	GRCm39 primary assembly
-gi	genome fasta index file	a fasta index file for GRCm39 primary assembly, obtained with "faidx" function from samtools
-r	genome segmentation file in GTF format	produced with iCount segment function from pre-filtered GENCODE vM28 primary assembly annotation
-s	length of the smoothing window	6
-p	percentile threshold for determination of thresholded crosslinks	0.7
-w	window around thresholded crosslinks for finding enriched kmers	20
-dw	distal window around enriched kmers to calculate relative k-mer occurrence	150
-n	number of top-ranked k-mers to cluster and visualise	20
-c	maximum number of k-mer clusters	5
-re	How to treat repeating regions within genome.	unmasked
-pos	percentile to set as threshold for relevant positions	None
-relax	Whether to relax automatically calculated threshold for relevant positions.	True
-sub	Whether to subsample crosslinks for motif enrichment, if there are many crosslinks in the dataset.	True
-a	Whether to save all outputs.	False

-seed	Set the seeds for random sampling to ensure reproducibility. For each random sample (n=100), the seed incrementally increases from 0 to 99.	True

6.1.3 k-mer logos and consensus sequences of PEKA k-mer groups

To visually represent k-mer groups, we used sequence logo representations. These were created by k-mer multiple-sequence alignment transformed to position-frequency matrix (PFM). To achieve this, pairwise sequence alignments of k-mers are first obtained by employing global Needleman-Wunsch algorithm with the *skbio.alignment* module v0.5.1 (Knight, n.d.), setting the score for a nucleotide match to 2 and the mismatch score to -1; other scoring parameters are left on their default settings—penalty for opening the gap is set to 5, the penalty for gap extension is 2, and terminal gaps in alignment are not penalised. Then, pairwise alignments are collated into the multiple-sequence alignment, starting with the highest scoring alignment (in case there are multiple alignments with the same score, the one that is the first by alphabetical sorting is taken) and aligning the second best pairwise alignment containing one of the motifs already included in the multiple-sequence alignment to it. This process is repeated with the next best scoring pairwise alignment until all k-mers are aligned in the multiple-sequence alignment. The multiple-sequence alignment is then transformed into a PFM, which denotes the frequency of each nucleotide at each position within the alignment. PFM is used to plot sequence logos with the *logomaker* (v0.8) library (Tareen & Kinney, 2019). By using a rolling window of a predefined length, a motif consensus can also be determined from the PFM by sliding the window across all PFM positions and summing the occurrences of nucleotides within a window. Where the sum is the greatest, the majority consensus sequence is derived from the PFM. In case of ties between two or more nucleotides, IUPAC nucleotide notation is used. In the case of multiple windows with the same highest scoring sum, the first window in the PFM to get that score is used to derive the consensus sequence. It should be noted that k-mer logos are not an accurate representation of the binding motifs, as PFMs are generated solely based on the sequence alignment of k-mers. Thus, k-mer logos do not necessarily reflect the precise relative positioning of the k-mers or their frequency in the foreground sequences that were used to identify the k-mers. Rather, k-mer logos are used to aid in the visualisation of the common sequence features of each investigated group of k-mers. Source code for k-mer clustering and for generation of k-mer logos is available on GitHub at https://github.com/ulelab/cluster_kmers and deposited to Zenodo (Kuret, 2023).

6.1.4 Metaprofile of average motif coverage around crosslinks

To visualise motif coverage around crosslink sites, we used the source code available on GitHub at https://github.com/ulelab/cv_coverage, which was also archived on Zenodo (A. G. Amalietti, 2021). This script visualises the mean k-mer coverage of a user-defined motif group around crosslink sites located within a specified transcriptomic region (corresponding to regions analysed by PEKA). The coverage can be computed on a full set of input crosslinks or on the subset of thresholded crosslinks (see the explanation of thresholding in the description of PEKA method). The algorithm first extracts the sequences, flanking the relevant crosslink sites. Then, the sequences are scanned with a rolling window equal to k-mer length to find parts of the sequence that match k-mers from the investigated motif group. All positions containing a motif from the investigated group

are given a score corresponding to the cDNA count of the evaluated crosslink position, and the remaining positions are scored 0. Scores at each position around crosslinks in the assessed region are summed across evaluated sequences and divided by the total cDNA count of evaluated crosslinks to generate the coverage showing the percent crosslink events overlapping with any k-mer from the group at each position. Optionally, the user can select to visualise the coverage unweighted by cDNA count, in which case all positions containing a motif from the investigated group are given a score of 1, remaining positions are scored 0; scores at each position around crosslinks in the assessed region are summed and divided by the number of evaluated crosslink sites to get the coverage, expressed as the percentage of crosslink positions that have a k-mer from the group aligned to a specific position relative to the crosslink site. Finally, coverage distributions are smoothed using a rolling mean, and the metaprofiles for the list of analysed crosslink files are plotted on the same graph.

For the analysis presented in Chapter 3 (Figure 3.4A,B), we calculated the average motif group coverage weighted by cDNA scores around tXn and oXn in the protein-coding gene region. The analysis was performed on the full set of crosslink sites; sequences were extracted in a −150...150nt window around the crosslinks; and the window of 6nt was used for smoothing. For the analysis presented in Chapter 4 (Figure 4.2B,D,C) we used crosslink sites in the 3′-UTR regions as input; sequences were extracted in a −300...300nt window around the crosslinks; the window of 20nt was used for smoothing. In all analyses, cDNA scores of input crosslinks were capped at 20, to avoid excessive contributions of few crosslink sites with disproportionately high cDNA scores.

6.2 Specificity of Protein-RNA Interactions Observed by CLIP

This section describes the details of methods used to produce the results described in Chapters 2 and 3. It covers data acquisition, data processing, and subsequent analyses. The text in this section is adapted from our published work (Kuret *et al.*, 2022).

6.2.1 Data acquisition and processing

6.2.1.1 CLIP experiments

To identify motifs enriched around RBP binding sites *in vivo*, we collected a large array of CLIP experiments for diverse RBPs and identified their crosslink positions on the transcriptome—either by registering cDNA truncations in eCLIP and iCLIP or T-C transitions in PAR-CLIP. Because different variants of CLIP methods apply different strategies to crosslinking and library generation, they require custom workflows for the determination of crosslink sites. Notably, while some differences in data processing steps stem from the differences in CLIP methods, others reflect the development and implementation of new bioinformatic approaches to optimise crosslink assignment. For example, iCLIP experiments used in this study were analysed on the older iMaps Genialis platform https://imaps.genialis.com/, which was developed until 2020. Conversely, pre-processed eCLIP reads were analysed with the *nf-core/clipseq* pipeline, which was released in 2021 (West et al., 2021). To ensure a comprehensive and accurate description of the data processing, we report the data acquisition and processing steps for each of the analysed CLIP variants in the following sections.

6.2.1.1.1 eCLIP

We downloaded the fastq files for eCLIP experiments from the ENCODE consortium (Davis et al., 2018; ENCODE Project Consortium, 2012). For the list of the files included in the study, see Additional file 2: Table S1 in our publication (Kuret *et al.*, 2022). Fastq files were processed to obtain the positions of cDNA truncations, which identify crosslink sites in eCLIP experiments. First, the reads were pre-processed with the *peka/eclip* pipeline, available at https://github.com/ulelab/peka-eclip, as follows: 3′-adapters were removed with Cutadapt v3.4 (M. Martin, 2011) using two rounds of adapter removal to account for double ligations in the ENCODE standard protocol (Van Nostrand et al., 2016); next, unique molecular identifiers were extracted and positioned at the end of the fastq header. To obtain the positions of crosslink sites, we discarded the forward sequencing reads, as only the reverse reads reliably identify cDNA truncations (Van Nostrand et al., 2016). Next, we applied the *nf-core/clipseq* pipeline (West et al., 2021) to further process the reads—applying Cutadapt quality trimming and filtering out reads that aligned to rRNA or tRNA with Bowtie 2 (Langmead et al., 2009)—and then align the remaining reads to the GRCh38 primary assembly human genome (using GENCODE V29 annotation) with STAR (Dobin et al., 2013). PCR duplicates were removed with UMI-tools (Smith et al., 2017), leveraging unique molecular identifiers. Finally, the crosslink positions were identified as the coordinate immediately 5′ to the alignment start, using BEDTools (Quinlan & Hall, 2010).

eCLIP narrowPeaks in GRCh38 genome build, combined from both replicates and filtered with corresponding SMInput control, were downloaded as BED files from the ENCODE consortium (Davis et al., 2018; ENCODE Project Consortium, 2012). For the list of peak files relevant for this study, see Additional file 2: Table S1 in our publication (Kuret *et al.*, 2022).

6.2.1.1.1.1 iCLIP

Crosslink positions from iCLIP experiments of hnRNPC (Zarnack et al., 2013), hnRNPL (Rossbach et al., 2014), TIA1 (Z. Wang et al., 2010) and TARDBP (Tollervey et al., 2011), were downloaded as bed files from the iMaps web platform for CLIP data analysis https://imaps.genialis.com/. The sequencing data from iCLIP experiments was processed on iMaps as follows: first, the single-end sequencing reads were trimmed with Cutadapt (M. Martin, 2011) to remove the 3′-adapters and nucleotides with high base calling uncertainty; reads were then filtered to a minimum length of 10 nucleotides and reads containing Ns were discarded; finally reads were mapped to the human GRCh38 genome build using STAR (Dobin et al., 2013) and PCR duplicates were removed with UMI-tools (Smith et al., 2017). Aligned reads were used to define the positions of crosslink sites with iCount v2.0.1 (Curk, 2019). SRA accession codes for all relevant iCLIP samples are listed in Additional file 3: Table S2 in our paper (Kuret *et al.*, 2022).

6.2.1.1.1.2 PAR-CLIP

For PAR-CLIP experiments analysed in this work, fastq files were downloaded from the SRA database; SRA accession codes for all relevant PAR-CLIP samples are listed in Additional file 3: Table S2 in our paper (Kuret *et al.*, 2022). First the 3′-adapter sequences were trimmed with Flexbar v2.5 (https://github.com/genome-vendor/flexbar) and the reads were collapsed to remove PCR duplicates. Then, the reads were sequentially mapped to reference transcripts by Bowtie 2 v2.3.2 (Langmead et al., 2009) by transferring the unmapped reads from the previous to the next mapping step; first, the reads were mapped to human pre-rRNA (GenBank U13369.1), followed by rRNA (GenBank NR_023363.1, NR_003285.2, NR_003287.2, NR_003286.2), snRNA, snoRNA, other ncRNAs (all from Ensembl, including RN7SL), tRNA (GtRNADb), mtDNA (GenBank AF347015.1), and

finally the human genome (GRCh38, primary assembly). The last genome-mapping step was performed by the STAR aligner v2.5.3a (Dobin et al., 2013), and only uniquely mapped reads were retained for further processing. Finally, crosslink sites were identified from T-C transitions, which we extracted using the SAMtools (H. Li et al., 2009) *mpileup* command and row_mpile_coverage_plus_TC.pl script (Schueler et al., 2014).

6.2.1.2 K-mer z-scores from in vitro experiments and mCross analysis of eCLIP data

To evaluate the ability of PEKA to enrich biologically relevant, RBP-specific motifs from CLIP data, we compared its motif discovery performance in CLIP to a related mCross method and to corresponding *in vitro* experiments. For this purpose, we obtained published 5-mer z-scores for 78 RNA-Bind-n-Seq datasets (Dominguez et al., 2018). The corresponding k-mer enrichment scores (R-scores) are accessible from the ENCODE resource, with the accession numbers and relevant concentrations specified in the supplementary material of the original publication (see Table S3 in (Dominguez et al., 2018)). The R-scores can be converted to z-scores using their mean and standard deviation. For the current study, the z-scores were obtained for the concentration of RBP that produced the highest motif enrichment. The z-scores for optimal RBP concentration were kindly provided to us in batch form by the papers' authors (Dominguez et al., 2018).

RNAcompete 7-mer z-scores were obtained from web supplementary data published by (Ray et al., 2013).

Raw and normalised mCross 7-mer z-scores for all ENCODE eCLIP datasets were kindly provided by the authors of the mCross paper (Feng et al., 2019).

7-mer z-scores were converted to 5-mer enrichment scores by calculating the arithmetic mean of z-scores across all 7-mers that contain a given 5-mer. 7-mers which contain a given 5-mer more than once were considered as many times as the number of instances of the contained 5mer. For illustration, when calculating the arithmetic mean of z-scores for a 5-mer "UUUUU," the 7-mer "UUUUUUG" would be considered two times ("[UUUUU]UG," "U[UUUUU]G"). For the ease of reproducing the findings of our study, we provided the 5-mer enrichment scores for RBNS, RNAC, and mCross datasets in supplementary materials of our paper (see Additional file 4: Table S3 in (Kuret *et al.*, 2022)).

6.2.1.3 Other data types

Enhanced RNA interactome capture (eRIC) data from Jurkat cells was obtained from Perez-Perri et al. (Perez-Perri et al., 2018); Supplementary Data 1. For our analyses, we visualised the log2-fold change in signal intensity in UV irradiated (UV+) over non-irradiated (UV-) samples.

The structural features—including domains, intrinsically disordered regions, and compositional biases—for all RBPs included in this study were downloaded from Uniprot as a single GFF file on May 22nd, 2022. We limited our query to reviewed Uniprot entries for human. Next, we filtered the GFF file to only include features termed "Domain", "Region", and "Compositional bias". We provided this detailed structural information as a supplement to our paper (see Additional file 10: Table S9 in (Kuret *et al.*, 2022)). For subsequent analyses, we labelled rare—occurring less than five times in all analysed RBPs—domains and compositional biases as 'other'.

6.2.2 Recall

Recall was calculated as a proportion of top 20 motifs from *in vitro* dataset (RBNS or RNAcompete) that are found among the top n motifs in the corresponding CLIP dataset; n is set to 50, unless specified otherwise, as in Figure 2.4A. Recall values for all analysed

CLIP datasets with available *in vitro* data are reported in our publication (Kuret *et al.*, 2022); see Additional file 8: Table S7. In cases where both RBNS and RNAcompete were available for a particular protein, we always prioritised RBNS over RNAcompete for the calculation of recall as RBNS z-scores were readily available for 5-mers, whereas RNAC required transformation from 7-mer to 5-mer scores. Moreover, RBNS performs enrichment at different protein concentrations, which increases its sensitivity.

6.2.1 Sequence-based clustering of k-mer groups

For sequence-based clustering of k-mers, individual motifs are first converted into tokens that reflect their sequence properties. Tokens resulting from a k-mer are all its subsequences, each combined with an end number denoting their cumulative incidence within a k-mer from left to right. For example, a k-mer "AGGU" is tokenised into "A1," "G1," "G2," "U1," "AG1," "GG1," "GU1," "AGG1," "GGU1," and "AGGU1." K-mer subsequences are marked with a number in the order in which they occur in the k-mer sequence. For example, two guanosines in the example k-mer produce two distinct tokens "G1" and "G2," one for each nucleotide. After tokenisation, pairwise Jaccard similarity is calculated for all k-mers in the group. Jaccard similarity is a quotient of the number of shared tokens between two compared motifs and the number of all tokens in the union formed by the k-mers. The k-mers are then clustered with an affinity propagation method, based on the resulting similarity matrix. Affinity propagation clustering was implemented with the *scikit-learn* v0.21 Python library, using the damping parameter of 0.5, the maximum allowed number of iterations set to 1000, and the number of convergent iterations set to 200. Affinity propagation clustering automatically determines the number of resulting clusters.

6.2.2 Clustering eCLIP datasets

For Figure 3.9A, we first split the eCLIP datasets into two groups based on whether or not they had available orthogonal in vitro data and then performed k-means clustering (implemented with the *scikit-learn* v0.21 Python library) on each group, using either an equally weighted combination of similarity index and recall, or just similarity index where no in vitro data was available. Prior to clustering, we normalised recall and similarity index across eCLIP datasets with min-max transformation to ensure an equal contribution of these parameters to clustering. eCLIP datasets with available in vitro data were split into 4 clusters, and datasets without available in vitro data were split into 3 clusters. For heatmap visualisation, we arranged clusters in each group by their median similarity index, and additionally, datasets within each cluster were arranged in ascending order based on their similarity index. eCLIP clusters and data related to the main heatmap in Figure 3.9A are available in our publication (Kuret *et al.*, 2022); see Additional file 7: Table S6.

6.2.3 Generation of differentially ranked motif groups between in vitro data and data produced by eCLIP or PAR-CLIP

The differentially enriched motif groups for Figure 3.4 were generated as follows. We obtained k-mers that were differentially enriched between in vitro approaches (RBNS or RNAC) and eCLIP motifs identified by PEKA, mCross or the local approach for all RBPs where data were available for at least one in vitro method and for mCross. For each k-mer, we compared its rank distribution across eCLIP datasets with the corresponding in vitro proteins and performed Welch's t-test to obtain a p-value. For each approach, we then extracted significantly differential k-mers that had a p-value < 0.01 and a fold change

greater than 1.5 or less than 0.66 (Figure 3.3B). Finally, differential k-mers were grouped based on the overlap between the analysis approaches and whether they were enriched or depleted in eCLIP, relative to in vitro data (Figure 3.3D,E).

To select an example RBP for each k-mer group shown in Figure 3.4A,B, we considered eCLIP datasets with available in vitro data. For each eCLIP dataset, we calculated its mean k-mer rank for the motif group in PEKA and in the corresponding in vitro data. Then, we calculated the mean value between PEKA and in vitro k-mer ranks and used this to find the RBP which had the highest (or second highest in the case of the AUAC group) enrichment of this motif group in both datasets (Figure 3.4A,B). In the case of group UUCG, we did not manage to find a dataset among those with available *in vitro* data that would enrich for these k-mers; therefore, we considered all eCLIP experiments and found DDX3X in both cell lines to be among the top three datasets for this motif group. In addition, a study reported that DDX3X binds in the vicinity of a motif composed of similar k-mers as in UUCG group (Calviello et al., 2021). Thus, we selected DDX3X in HepG2 cells as an example dataset for this motif group (Figure 3.4A).

6.2.4 STREME

We ran STREME (v5.4.1) on sequences extracted from Clippy peaks and narrowPeaks for 10 eCLIP datasets, shown in Figure 3.7. Sequences were encoded in standard RNA alphabet, the width of motifs (w) was fixed at 5 nucleotides, for easier comparison with 5-mers, produced by PEKA. The objective function to optimise for motif discovery was set to differential enrichment (default), the number of seeds to evaluate for each width was set to 100, patience was set to 10, and n-order of shuffle was set to 2 (default). STREME stopped after 10 consecutive motifs exceeded the p-value threshold (0.05). The command is written below:

```
streme --p fasta_file --w 5 --objfun de --neval 100 --patience 10 --order 2
-rna
```

To convert STREME motifs to k-mers, we decoded their consensus sequences based on the IUPAC nucleotide code. For example, the motif "UCWUC" was to be converted into two k-mers "UCAUC" and "UCUUC". Only the k-mer which ranked most highly in the *in vitro* data was used to represent the STREME motif.

6.3 Specificity of LIN28A-Mediated mRNA Decay in Early Embryonic Development

6.3.1 Data collection

3′-end sequencing experiments and iCLIP data related to the study of LIN28A-mediated RNA decay were produced by Miha Modic, and are described in the manuscript, which is currently under review and available as a preprint (Modic et al., 2023). As this work focuses on computational analysis of these data, we will cover these aspects in detail; however, for details related to experimental design, cell line generation, and experimental protocols, we refer the readers to the preprint (Modic et al., 2023).

All experiments analysed in this study were performed on inducible mouse embryonic stem cell (mESC) lines, in which the *Lin28a* gene was knocked out using the CRISPR/Cas9 genome editing technology, and then introduced as a transgene in the doxycycline inducible

PiggyBac system. The experiments analysed in this work were performed on cell lines expressing one of these transgenes:

- FLAG-tagged WT LIN28A protein, referred to as LIN28A-WT,
- FLAG-tagged phosphomutant LIN28A in which the serine at position 200 is replaced with alanine, referred to as LIN28A-S200A.

Naïve mESCs were maintained by culturing them in the medium, which contains LIF, a ligand of the Jak–Stat pathway, in combination with two inhibitors of the kinases GSK3 and MEK, i.e., the 2iLIF conditions (Betto et al., 2021). To trigger the exit of cells from naïve cell state, the cells were harvested and transferred the 'priming medium', in which the MEK inhibitor is removed, and instead basic fibroblast growth factor (bFGF or FGF2) is added, which activates the MEK/ERK signalling pathway (Fathi et al., 2017). 3′-end sequencing experiments and iCLIP experiments performed in the state of MEK activation were performed 6 hours after the medium change.

6.3.1.1 3′-end sequencing experiments

In this study, we analysed 3′-end RNA sequencing experiments performed on clones in which the native *lin28a* gene was knocked out and that instead successfully expressed either the LIN28A-WT or the LIN28A-S200A transgene. All sequencing experiments considered here were performed in 6 hours after MEK/ERK activation in the following conditions: (i) the expression of transgene LIN28A was not induced, i.e., the knockout condition; (ii) the expression of transgene LIN28A-WT was induced, i.e., the LIN28A-WT condition; (iii) the expression of transgene LIN28A-S200A was induced, i.e., the LIN28A-S200A condition. These experiments enabled us to analyse changes in gene expression, with respect to LIN28A expression and phosphorylation. Sequenced reads were processed as described in the 'RNA-seq processing and analysis' section. The experiments analysed in this study are listed in Table 6.3.

Table 6.3: 3′-end RNA-seq experiments analysed in this study.

Name	**Condition**	**N replicates**	**SRA Accessions**
KO	+MEK; KO	3	ERX10678018 ERX10678019 ERX10678020
LIN28A-WT	+MEK; iLIN28A-WT	3	ERX10678033 ERX10678034 ERX10678037
LIN28A-S200A	+MEK; iLIN28A-S200A	5	ERX10678024 ERX10678025 ERX10678026 ERX10678029 ERX10678030
Naïve	-MEK; iLIN28A-WT	4	ERX10678031 ERX10678032 ERX10678035 ERX10678036

6.3.1.2 iCLIP

iCLIP experiments analysed in this study are deposited in the ENA database under the accession number PRJEB60519 and to flow.bio webserver (see collection information in Table 6.4). Individual experiments are listed in Table 6.5, together with their ENA sample accession code and sample name in flow.bio. The experiments were processed and analysed as described in the 'iCLIP data processing and analysis' section. We excluded the second replicate of LIN28A-S200A from the analysis, due to low sequencing depth. Accordingly, we did not perform quantitative analyses of the LIN28A-S200A iCLIPs and only performed qualitative analyses in the context of LIN28A-WT -MEK experiments, in which the binding of unphosphorylated protein is interrogated.

Table 6.4: Collections of processed iCLIP data on flow.bio.

Collection Name	**Owner**	**Contents**	**Link**
LIN28A_signalling_repeated_correct	Miha Modic	iCLIP experiments targeting LIN28A, listed in Table 6.5.	https://app.flow.bio/projects/882635250203/
PABPC1PABPC4	Miha Modic	iCLIP experiments targeting PABPC1 and PABPC4 in LIN28A KO cells, with or without the induction of LIN28A transgene.	https://app.flow.bio/projects/340215254997/

Table 6.5: iCLIP experiments analysed in this study.

IP target	**Condition**	**Replicate**	**ENA Sample Accession**	**Sample Name (flow.bio)**
LIN28A-WT	-MEK; iLIN28A-WT	1	SAMEA112855867	LIN28A-WT_ESCiLIF0220626_MM_1
		2	SAMEA112855868	LIN28A-WT_ESCiLIF0220626_MM_2
		3	SAMEA112855869	LIN28A-WT_ESCiLIF-OLD0220626
LIN28A-WT	+MEK; iLIN28A-WT	1	SAMEA112855865	LIN28A-WT_ESC_LIF-CHIR-FGF0220626_MM_1
		2	SAMEA112855866	LIN28A-WT_ESC_LIF-CHIR-FGF0220626_MM_2
LIN28A-S200A	+MEK; iLIN28A-S200A	1	SAMEA112855864	LIN28A-S200A_ESC_LIF-CHIR-FGF0220626_MM_1
PABPC1	+MEK; iLIN28A-WT	1	SAMEA112855856	DOX_C1_Crick1
		2	SAMEA112855857	DOX_C1_Crick2
PABPC1	+MEK; KO	1	SAMEA112855858	KO_C1_Crick1

		2	SAMEA112855859	KO_C1_Crick2
PABPC4	+MEK; iLIN28A-WT	1	SAMEA112855860	DOX_C4_Proteintech_1
		2	SAMEA112855861	DOX_C4_Proteintech_2
PABPC4	+MEK; KO	1	SAMEA112855862	KO_C4_Proteintech_1
		2	SAMEA112855863	KO_C4_Proteintech_2
PABPC1	-MEK; mESCs expressing native LIN28A (WT)	Combined replicates 1 and 2	This sample is available at Zenodo; DOI: 10.5281/zenodo.10054232	PABPC1_ESC_WT_grouped

6.3.2 RNA-seq processing and analysis

3′-end sequencing reads were quantified using Salmon (Patro et al., 2017), as described in (Corley et al., 2019); the transcriptome was built from GENCODE M22 transcripts (the GRCm38 genome build). This provided us with the table of transcript-per-million (TPM) values, indicative of transcript expression levels in the sample. To find genes that were differentially regulated upon phosphorylation of LIN28A, we conducted differential expression analysis comparing LIN28A-KO cells without the induction of a transgene, to cells in which either LIN28A-WT or LIN28A-S200A was induced; all cells were treated with FGF2 for 6h before sequencing to activate MEK/ERK signalling.

Differential expression analysis was performed with DESeq2 (Love et al., 2014), applying the effect-size shrinkage with the *apeglm* package (A. Zhu et al., 2019). TPM values for the full transcriptome were imported with the *tximport* package and converted to gene-level counts. These values were used to construct the DESeq data set and normalised with the *estimateSizeFactors* function. Then, lowly expressed genes were filtered out, such that only the genes in which at least two replicates had a normalised count greater than 5 were used as input to differential analysis. Differentially expressed genes were obtained by applying criteria for adjusted p-value < 0.05 and fold-change ≥ 1.5 or fold-change ≤ 0.66; The alpha value for false-discovery-rate was set to 0.1. The control group of genes was defined by adjusted p-value ≥ 0.05 and $|\log2(\text{fold-change})| < 0.5$. For the subsequent analysis of 3′-UTR features, we filtered the gene groups to contain only protein-coding genes with minimum 3′-UTR length of 100 nts and an expression level ≥ 5 TPM in either LIN28A-KO or LIN28A-WT in the FGF2-treated condition. This resulted in 1183 downregulated genes, 989 upregulated genes and 2703 control genes.

For subsequent analyses we focused on exons, particularly the 3′-UTRs. To determine the relevant exons, we defined one 'representative transcript' for each gene; the relevant transcript was annotated based on the most abundant mRNA isoform in naïve ESCs cells expressing LIN28A-WT cultured in 2iL medium (Naïve, Table 6.3). Expression levels were evaluated from TPM values obtained with Salmon, by taking a mean TPM value for each transcript across all replicate experiments.

6.3.3 iCLIP data processing and analysis

6.3.3.1 Processing to obtain crosslink sites

iCLIP reads for this study were analysed on the iMaps webserver (https://imaps.goodwright.com/), which has since migrated to flow.bio (Capitanchik et al., 2023). First, reads were demultiplexed using Ultraplex v1.2.5 (Wilkins, 2021) and

barcodes were removed using Cutadapt v3.4 (M. Martin, 2011). Reads were then premapped to mouse genome build (GRCm39 GENCODE M28 annotation) with Bowtie (Langmead et al., 2009) and then aligned with STAR v2.7.9a (Dobin et al., 2013), followed by removal of PCR-duplicates using UMI-tools (Smith et al., 2017) and identification of crosslink-events, at the nucleotide preceding each sequenced read. Specific settings to the pipeline's software and all files generated during data processing are available from the flow.bio interactive webserver view (see Table 6.4 for collection links and Table 6.5 for sample names).

6.3.3.2 Peak-calling and motif analysis

Peaks of CLIP signal were identified with Clippy (Kuret *et al.*, 2022), and used together with the crosslink sites to run PEKA v1.0.0 as described in the 'General' section of Methods. Settings for PEKA are listed in Table 6.2. For Clippy and PEKA, the GENCODE primary assembly annotation M28 was filtered to retain only entries with transcript support level 1 or 2 in genes where such transcripts were available and used to produce a segmentation file with the *get_segments* function from the iCount tool (Curk, 2019).

6.3.3.3 Analysis of crosslink proportions in exons

For the analysis of crosslink proportions in different types of exonic regions, we extracted the 5′-UTR, CDS, and 3′-UTR regions of representative transcripts in DOWN, Control, and UP groups of genes. We merged replicate iCLIP experiments by summing up cDNA counts at overlapping positions and then computed the total number of cDNAs (from merged replicates) that fell into a particular exonic region in each transcript. We converted these counts into percentages to ascertain the absolute proportion of crosslinks that occur in the 5′-UTR, CDS, and 3′-UTR regions; this proportion was then normalised by region length to 1000nt to get a measure of crosslink density in each exon type expressed as % cDNA per 1000nt. The distribution of these densities across UP, DOWN, and Control genes are shown in Figure 4.1E and Figure 4.4A.

6.3.3.4 Identification of motif groups from CLIP data

For identification of enriched motif groups, we obtained PEKA results in 3′-UTRs (files ending in *5mer_distribution_UTR3.tsv; accessible from flow.bio see Table 6.4 for link to collection) for all LIN28A-WT (in 2iL and FGF2-treated cells) and LIN28A-S200A (in FGF2-treated cells) iCLIPs, except for LIN28A-S200A_ESC_LIF-CHIR-FGF0220626_MM_2, which was excluded from subsequent analyses due to low read coverage.

First, we combined enriched k-mers (p-value < 0.05 in PEKA result files) from all samples into one group of unique k-mers (n=93), which we then clustered based on their sequence similarity and their ranking in PEKA. To achieve this, we computed Euclidean distances between k-mer ranks in PEKA and Jaccard distances between k-mer sequences. To obtain sequence distances, each k-mer sequence was converted into a list of all possible substrings with length less than k, for example, a trimer 'UGA' would be converted into 'U', 'G', 'A', 'UG' and 'GA'. Then, Jaccard similarity was calculated on sets of substrings for each pair of k-mers. Jaccard similarity is a quotient of the number of shared substrings between two k-mers and the number of all substrings in the union. Finally, Jaccard similarities were converted into distances by subtracting the similarity values from 1. Next, we applied standard scaling to each of the resulting distance matrices to account for differences in variance between the two metrics, followed by min-max scaling. Normalised matrices were combined with Pythagorean addition and used to cluster k-mers using

scipy.hierarchy (Virtanen *et al.*, 2020), with *correlation* to compute distances and UPGMA algorithm to perform clustering. Finally, the dendrogram was plotted with *scipy.hierarchy.dendrogram* and the tree was cut into 4 clusters (Figure 4.2A) to obtain the following motif groups: the AGGG-motif group, representative of the ZnF domain; the UGGG-motifs; the GAU-motif group, bound by CDS; and the AUU-motif group. Finally, we combined AGGG and UGGG into one motif group (WGG), as they exhibited similar k-mer sequences and enrichment patterns in PEKA (Figure 4.2A). K-mer logos for resulting motif groups were plotted as described in the '6.1.3 k-mer logos and consensus sequences of PEKA k-mer groups' section. The source code for k-mer clustering and for generation of k-mer logos is available on GitHub and deposited to Zenodo (Kuret, 2023).

6.3.3.5 Proportional crosslinking distributions

To produce the metaprofiles of iCLIP signal relative distribution across protein-coding exons, we first obtained the 5′-UTR, CDS, and 3′-UTR exons of UP/Control and DOWN genes, corresponding to the representative transcript for each gene. To account for splicing, exons from the specific region were concatenated from the 5′ to the 3′-end. Then, we obtained the per-nucleotide raw crosslink coverage (i.e. cDNA counts) in these regions with *bedtools coverage* utility (Quinlan & Hall, 2010). This quantification was performed for the following merged iCLIP samples: LIN28A-WT in the FGF2-treated condition (+MEK); LIN28A-WT in the 2iLIF-treated condition (-MEK); PABPC1 and PABPC4 in LIN28A knockout cells, treated with FGF2, with or without the induction of LIN28A-WT (Table 6.5).

For each quantified exonic region in a specific transcript, the cDNA counts of individual iCLIP experiments were binned into 100 equal-sized bins, based on their position within the region. Then, we computed the percentage of crosslinks in each bin, for every quantified region. To evaluate the global binding patterns of LIN28A and PABPC proteins, we then calculated the mean of crosslink percentages in each bin, across all exons of the same type (e.g., CDS). The mean percentages of cDNAs in the region were evaluated separately for UP/Control and DOWN gene groups.

The bin means were scaled with *RobustScaler* module from the *scikit-learn* library (Pedregosa et al., 2011), to correct for outliers and to identify regions within the exons where the iCLIP signal was the strongest. After scaling, the scaled mean values were smoothed using a window of 5 bins, and the smoothed data was plotted as a line graph to visualise the distribution of crosslink counts across regions (Figure 4.3A, Figure 4.4B).

6.3.3.6 Estimation of gene-level expression

Expression levels for each gene were obtained by summing up TPM (transcript-per-million) values of transcripts with the same stable Ensembl gene id and then averaging these sums across replicate 3′end sequencing experiments. Transcript TPM values were obtained from 3′end sequencing data with Salmon, as described in "RNA-seq processing and analysis". These gene-level expression estimates are referred to in the following text as gene-level TPM.

6.3.3.7 Metaprofiles of normalised crosslink coverage

To generate the metaprofiles of library- and expression-normalised crosslink coverage upstream around canonical PAS (Figure 4.3B, Figure 4.4C) and around peak centres (Appendix A.1-B,C), we first computed the per-nucleotide raw crosslink coverage (i.e., cDNA counts) for individual iCLIP samples in regions of interest with *bedtools coverage* utility (Quinlan & Hall, 2010). Next, raw crosslink coverage was normalised by sequencing

depth for each iCLIP, to CPM (crosslinks-per-million) values; this was followed by normalisation within each gene to its expression level in a relevant experimental condition, using gene-level TPM values, obtained as described above.

This yielded a set of quantified genomic regions with crosslink coverage at each position expressed as CPM-per-TPM. Then, we smoothed the expression-normalised coverage within each quantified region, using a rolling mean across 20nt and a triangular window. Smoothed value was assigned to the position located at the centre of the window. Missing values resulting from smoothing at the 5′- and 3′-ends of the region were replaced with the closest valid observation. Finally, we computed the mean of smoothed coverages across all regions of interest, and a 95% confidence interval at each position within the region.

For metaprofiles around PAS (Figure 4.3B, Figure 4.4C), the 3′-UTRs with the canonical PAS sequence 'AATAAA' and a length of minimum 300 nts were included in the analysis. In case of multiple canonical PAS within a 3′-UTR, only the PAS closest to the 3′-UTR termini was used. This included 1739 PAS from Control, 831 from DOWN and 633 from UP genes.

For the metaprofiles shown in Appendix A.1-C we defined the regions as 100nts upstream and downstream from the centres of PABPC peaks, that were located either within the last 200nts of the 3′-UTRs or elsewhere in the 3′-UTR. PABPC peaks were obtained by merging book-ended or overlapping Clippy peaks of PABPC1 or PABPC4 iCLIPs in LIN28A KO cells with and without induced LIN28A expression, using *bedtools merge* (Quinlan & Hall, 2010). For the metaprofiles shown in Appendix A.1-B we defined the regions as 100nts upstream and downstream from the centres of LIN28A peaks located in the 3′-UTRs that overlap or do not overlap with PABPC peaks. A minimum overlap of 1nt was used and peak overlap was identified with *bedtools intersect* (Quinlan & Hall, 2010). LIN28A peaks were obtained by merging book-ended or overlapping Clippy peaks of LIN28A-WT (in 2iL and FGF2-treated cells) and LIN28A-S200A (in FGF2-treated cells), using *bedtools merge*; PABPC peaks were obtained by merging PABPC1 and PABPC4 peaks from LIN28A KO cells with and without induced LIN28A expression.

6.3.3.8 Visualisation of iCLIP data in 3′-UTRs of naïve genes

6.3.3.8.1 Comparison of LIN28A binding to 3′-UTR termini of naïve genes before and after MEK/ERK activation

Barplot in Figure 4.3E shows the log2 fold-change in expression-normalised crosslink coverage (expressed as CPM-per-TPM values) between merged iCLIPs of LIN28A-WT in FCL-treated condition (+MEK) and LIN28A-WT in 2iL-treated condition (-MEK) in genes of naïve regulon. Naïve genes included in this analysis were filtered to have a minimum 3′-UTR length of 500 nts and sufficient expression in level ($\geq$ 5 TPM in either LIN28A-KO or LIN28A-WT in the FGF2-treated condition). This resulted in the inclusion of 16 out of the 22 genes of naïve regulon.

To produce the plot in Figure 4.3E, we first computed raw crosslink coverage for the relevant iCLIPs in 500nt regions upstream of their annotated 3′-UTR termini (according to the representative transcript). Then, we summed raw crosslinking signal in bins of 20nts and converted the crosslink counts into percentage of counts within each bin, for each quantified region. The value of 1 was then added to each bin to avoid division by zero. We proceeded to calculate a log2 fold-change between the percentage of crosslinks in FGF2-treated condition relative to 2iLIF-treated condition for each bin, in each naïve gene. The mean log2 fold-change in each bin, obtained across all evaluated genes, was indicated in the bar plot, with error-bars indicating 95% confidence intervals.

6.3.3.8.2 Crosslinking profiles of LIN28A and cytoplasmic PABPs across 3′-UTRs of three naïve genes

Visualisation of iCLIP signal within 3′-UTRs of Tfcp2l1, Zfp281, and Esrrb (Appendix A.1-A,D,E) was performed with a *clipplotr* tool (Chakrabarti *et al.*, 2021), using normalisation of iCLIP cDNA counts at a given crosslink site by the experimental library size. Normalised counts were smoothed with a rolling mean across a window of 50 nt and plotted across the regions of interest. Auxiliary tracks below the crosslinking signal represent LIN28A binding sites corresponding to WGG-, GAU- and WUU- motif groups (see '6.3.3.9 Motif-based binding site assignment' section for details), the location of AAA trimer and the location of canonical PAS 'AAUAAA'.

6.3.3.9 Motif-based binding site assignment

We defined LIN28A binding sites corresponding to WGG-, GAU-, and WUU-motif groups using the approach of motif-based binding site assignment, which was adopted from the approach used previously by Hallegger et al. (Hallegger *et al.*, 2021). First, crosslink sites from all LIN28A-WT (in 2iL and FGF2-treated cells) and LIN28A-S200A (in FGF2-treated cells) iCLIP samples, except for LIN28A-S200A_ESC_LIF-CHIR-FGF0220626_MM_2, were merged by summing up cDNA counts at overlapping positions. Then, binding sites were assigned separately for each motif group where k-mers from a given group were located at relevant positions in the range of ±20nt around crosslink sites. Relevant positions for each k-mer were determined by combining relevant positions identified by PEKA (*prtxn*, available in files ending in *5mer_distribution_UTR3.tsv on flow.bio; see Table 6.4 for links) from all LIN28A iCLIP samples. The script for motif-based binding site assignment is available at GitHub (A. Amalietti, 2021). Motif-based binding sites of LIN28A are shown as auxiliary tracks in Appendix A.1-A,D,E.

6.3.4 Modelling of protein structure

For Figure 4.2B,C,D, we modelled a crystal structure of LIN28A in complex with let-7d microRNA pre-element (PDB ID: 3trz, (Nam *et al.*, 2011)) using ChimeraX (Pettersen *et al.*, 2021). The protein is displayed in light-green colour and the RNA chain is displayed in grey. Nucleotides interacting with the protein are represented with sticks and colour-coded; adenosines are shown in green, guanosines in yellow and uridines in red. Hydrogen bonds between the protein and the coloured RNA motifs are shown as dashed blue lines. Planes of aromatic rings with a potential to form π-π stacking interactions are shown with mesh.

6.3.5 Trimer valency and motif coverage in 3′-UTR regions

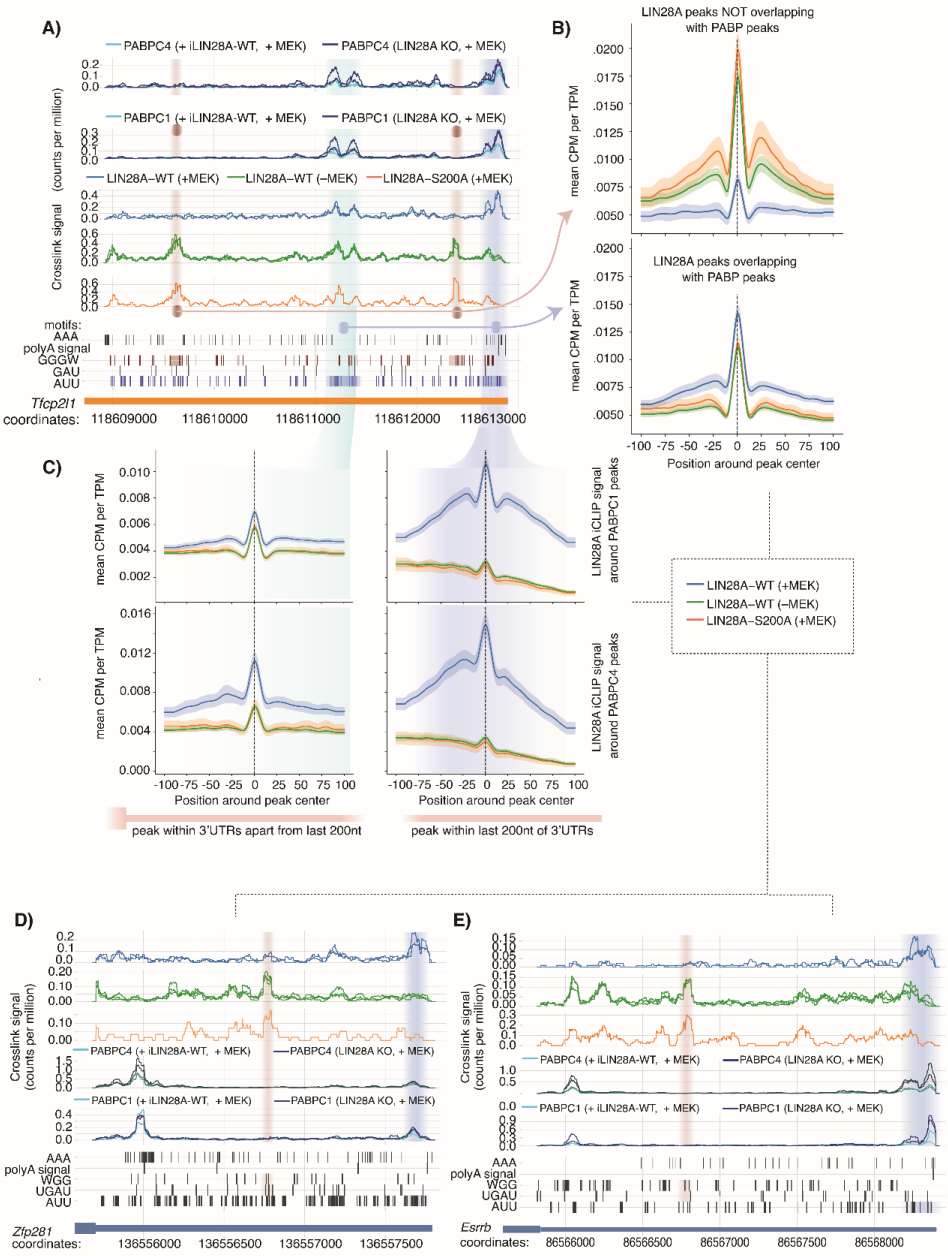
To assess the number of trimers in regions 100 nts upstream of PAS for genes belonging to Control, DOWN and UP groups (Figure 4.3D, Figure 4.4D), we obtained genomic sequences of those regions using *bedtools getfasta* (Quinlan & Hall, 2010). For each gene, only the PAS closest to the 3′-UTR termini with canonical 'AATAAA' sequence was used, which resulted in inclusion of 2126 PAS in Control, 1002 in DOWN and 778 in UP group. Next, we located the motifs of interest in these regions with *seqkit locate* v2.3.1 and wrote their genomic coordinates into a bed file. Then, we counted the number of relevant trimers in individual 3′-UTRs and plotted the distribution of counts for Control, DOWN and UP genes.

To compute the density of motif coverage, shown in Figure 4.3C, we obtained genomic sequences of relevant regions using *pybedtools.sequence* v0.9.0. Next, we scanned these

sequences with a window of length five and checked whether the sequence in a window corresponded to any of the 5-mers in the given motif group (namely the WUU-, WGG- or GAU-motif group, see '6.3.3.4 Identification of motif groups from CLIP data'). If the sequence in a window matched with a 5mer in a motif group, the nucleotides in the window received a score of 1. Next, we binned the regions into bins of 10nt and computed the % of covered nts in each bin. Finally, we computed the average bin coverage across all evaluated regions and plotted these values in the form of a heatmap. Included 3′-UTRs in each heatmap correspond to those shown in the associated metaprofiles; for selection criteria see '6.3.3.7 Metaprofiles of normalised crosslink coverage'.

Appendix A

A.1 Supplementary Figure to Chapter 4



The binding of LIN28A and cytoplasmic PABPs on 3′-UTRs of naïve mRNAs.

Continued on next page.

Continued from previous page.
(A) Library-normalised crosslink profiles of ~5kb part of 3′-UTR of naïve pluripotency factor Tfcp2l1 for LIN28A-WT (in 2iL and FGF2-treated cells), LIN28A-S200A (FGF2-treated cells) iCLIPs and PABPC1/4 iCLIPs (LIN28A KO with and without induced LIN28A expression). Auxiliary tracks below show motif-based binding sites of LIN28A that correspond to WGG-, GAU- and AUU-motif groups (Methods), to AAA trimer, and to canonical 'AAUAAA' PAS. **(B)** Line plot shows metaprofiles of expression-normalised crosslink coverage for LIN28A iCLIPs around the centres of LIN28A peaks in 3′-UTRs, merged together from all samples (Methods). Peaks were stratified into those that overlap with peaks of PABPC1/4 (bottom) and those that do not (top). Metaprofile shows that the crosslinking signal of phosphorylated LIN28A-WT increases around peaks that overlap with PABPC1/4 binding and decreases around peaks that do not overlap with PABPC1/4 peaks. Shaded areas indicate a 95% confidence interval. **(C)** Line plot shows metaprofiles of expression-normalised crosslink coverage for LIN28A iCLIPs around the centres of PABPC1 peaks (above) and PABPC4 peaks (below) located within the final 200nts of 3′-UTRs (right) or the rest of the 3′-UTRs (left). Metaprofile shows a prominent increase in crosslink signal of pLIN28A around PABP peaks located in the last 200nt and thus indicates phosphorylation-induced repositioning of LIN28A to the 3′-termini. Shaded areas indicate a 95% confidence interval. More information in Methods. **(D,E)** Library-normalised crosslink profiles of naïve pluripotency factors Zfp281 (D) and Esrrb (E) 3′-UTR for LIN28A-WT (in 2iL and FGF2-treated cells), LIN28A-S200A (FGF2-treated cells) iCLIPs and PABPC1/4 iCLIPs (LIN28A KO with and without induced LIN28A expression). Auxiliary tracks below show motif-based binding sites of LIN28A that correspond to WGG-, GAU- and AUU-motif groups (Methods), to AAA trimer, and to PAS.

References

Agostini, F., Cirillo, D., Ponti, R. D., & Tartaglia, G. G. (2014). SeAMotE: a method for high-throughput motif discovery in nucleic acid sequences. *BMC Genomics*, *15*, 925.

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

Altieri, J. A. C., & Hertel, K. J. (2021). The influence of 4-thiouridine labeling on pre-mRNA splicing outcomes. *PloS One*, *16*(12), e0257503.

Amalietti, A. (2021). *ulelab/bs_assign: v0.0.0*. https://doi.org/10.5281/zenodo.8388765

Amalietti, A. G. (2021). *Comparative Visualisation of Average Motif Coverage*. Zenodo. https://doi.org/10.5281/ZENODO.8386510

Amalietti, A. G., Kuret, K., & Ule, J. (2022). *PEKA - Positionally-enriched k-mer analysis v0.1.6 (Zenodo)* (v0.1.6) [Computer software]. https://doi.org/10.5281/zenodo.6984815

Ambros, V., & Horvitz, H. R. (1984). Heterochronic mutants of the nematode Caenorhabditis elegans. *Science*, *226*(4673), 409–416.

Ascano, M., Hafner, M., Cekan, P., Gerstberger, S., & Tuschl, T. (2012). Identification of RNA-protein interaction networks using PAR-CLIP. *Wiley Interdisciplinary Reviews. RNA*, *3*(2), 159–177.

Auweter, S. D., Oberstrass, F. C., & Allain, F. H.-T. (2006). Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Research*, *34*(17), 4943–4959.

Bahrami-Samani, E., Penalva, L. O. F., Smith, A. D., & Uren, P. J. (2015). Leveraging cross-link modification events in CLIP-seq for motif discovery. *Nucleic Acids Research*, *43*(1), 95–103.

Bailey, T. L. (2021). STREME: Accurate and versatile sequence motif discovery. *Bioinformatics* . https://doi.org/10.1093/bioinformatics/btab203

Bailey, T. L., Williams, N., Misleh, C., & Li, W. W. (2006). MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Research*, *34*(Web Server issue), W369-73.

Balachandran, S., & Narendran, A. (2023). The Developmental Origins of Cancer: A Review of the Genes Expressed in Embryonic Cells with Implications for Tumorigenesis. *Genes*, *14*(3). https://doi.org/10.3390/genes14030604

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

Balzer, E., & Moss, E. G. (2007). Localization of the developmental timing regulator Lin28 to mRNP complexes, P-bodies and stress granules. *RNA Biology*, *4*(1), 16–25.

Betto, R. M., Diamante, L., Perrera, V., Audano, M., Rapelli, S., Lauria, A., Incarnato, D., Arboit, M., Pedretti, S., Rigoni, G., Guerineau, V., Touboul, D., Stirparo, G. G., Lohoff, T., Boroviak, T., Grumati, P., Soriano, M. E., Nichols, J., Mitro, N., …

Martello, G. (2021). Metabolic control of DNA methylation in naive pluripotent cells. *Nature Genetics*, *53*(2), 215–229.

Bischler, T. (2017). *PEAKachu: Peak calling tool for CLIP-seq data* (v0.1.0) [Computer software]. Github. https://github.com/tbischler/PEAKachu

Boyle, E. A., Her, H.-L., Mueller, J. R., Naritomi, J. T., Nguyen, G. G., & Yeo, G. W. (2023). Skipper analysis of eCLIP datasets enables sensitive detection of constrained translation factor binding sites. *Cell Genomics*, 100317.

Broughton, K., Esquer, C., Echeagaray, O., Firouzi, F., Shain, G., Ebeid, D., Monsanto, M., Yaareb, D., Golgolab, L., Gude, N., & Sussman, M. A. (2022). Surface Lin28A expression consistent with cellular stress parallels indicators of senescence. *Cardiovascular Research*. https://doi.org/10.1093/cvr/cvac122

Buchbender, A., Mutter, H., Sutandy, F. X. R., Körtel, N., Hänel, H., Busch, A., Ebersberger, S., & König, J. (2020). Improved library preparation with the new iCLIP2 protocol. *Methods (San Diego, Calif.)*, *178*, 33–48.

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

Burger, K., Mühl, B., Kellner, M., Rohrmoser, M., Gruber-Eber, A., Windhager, L., Friedel, C. C., Dölken, L., & Eick, D. (2013). 4-thiouridine inhibits rRNA synthesis and causes a nucleolar stress response. *RNA Biology*, *10*(10), 1623–1630.

Calviello, L., Venkataramanan, S., Rogowski, K. J., Wyler, E., Wilkins, K., Tejura, M., Thai, B., Krol, J., Filipowicz, W., Landthaler, M., & Floor, S. N. (2021). DDX3 depletion represses translation of mRNAs with complex 5′ UTRs. *Nucleic Acids Research*, *49*(9), 5336–5350.

Capitanchik, C., Ireland, S., Harston, A., Cheshire, C., Marc Jones, D., Lee, F. C. Y., de los Mozos, I. R., Iosub, I. A., Kuret, K., Faraway, R., Wilkins, O. G., Arora, R., Hallegger, M., Modic, M., Chakrabarti, A. M., Luscombe, N. M., & Ule, J. (2023). Flow: a web platform and open database to analyse, store, curate and share bioinformatics data at scale. In *bioRxiv* (p. 2023.08.22.544179). https://doi.org/10.1101/2023.08.22.544179

Capitanchik, C., Jones, M., Ule, J., & M. Luscombe, N. (2022, July 17). *Clippy peak caller v1.5.0*. GitHub. https://github.com/ulelab/clippy

Chakrabarti, A. M., Capitanchik, C., Ule, J., & Luscombe, N. M. (2021). clipplotr - a comparative visualisation and analysis tool for CLIP data. In *bioRxiv* (p. 2021.09.10.459763). https://doi.org/10.1101/2021.09.10.459763

Chakrabarti, A. M., Haberman, N., Praznik, A., Luscombe, N. M., & Ule, J. (2018). Data Science Issues in Studying Protein-RNA Interactions with CLIP Technologies. *Annual Review of Biomedical Data Science*, *1*(1), 235–261.

Chang, M.-Y., Oh, B., Choi, J.-E., Sulistio, Y. A., Woo, H.-J., Jo, A., Kim, J., Kim, E.-H., Kim, S. W., Hwang, J., Park, J., Song, J.-J., Kwon, O.-C., Henry Kim, H., Kim, Y.-H., Ko, J. Y., Heo, J. Y., Lee, M. J., Lee, M., … Lee, S.-H. (2019). LIN28A loss of function is associated with Parkinson's disease pathogenesis. *The EMBO Journal*, *38*(24), e101196.

Chen, C. Y., & Shyu, A. B. (1995). AU-rich elements: characterization and importance in mRNA degradation. *Trends in Biochemical Sciences*, *20*(11), 465–470.

Chen, X., Castro, S. A., Liu, Q., Hu, W., & Zhang, S. (2019). Practical considerations on performing and analyzing CLIP-seq experiments to identify transcriptomic-wide RNA-protein interactions. *Methods* , *155*, 49–57.

Chi, S. W., Zang, J. B., Mele, A., & Darnell, R. B. (2009). Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, *460*(7254), 479–486.

Cho, J., Chang, H., Kwon, S. C., Kim, B., Kim, Y., Choe, J., Ha, M., Kim, Y. K., & Kim, V. N. (2012). LIN28A is a suppressor of ER-associated translation in embryonic stem cells. *Cell*, *151*(4), 765–777.

Conn, S. J., Pillman, K. A., Toubia, J., Conn, V. M., Salmanidis, M., Phillips, C. A., Roslan, S., Schreiber, A. W., Gregory, P. A., & Goodall, G. J. (2015). The RNA binding protein quaking regulates formation of circRNAs. *Cell*, *160*(6), 1125–1134.

Corley, S. M., Troy, N. M., Bosco, A., & Wilkins, M. R. (2019). QuantSeq. 3' Sequencing combined with Salmon provides a fast, reliable approach for high throughput RNA expression analysis. *Scientific Reports*, *9*(1), 18895.

Curk, T. (2019). *iCount: iCount, protein-RNA interaction analytics.* Github. https://github.com/tomazc/iCount

Danan, C., Manickavel, S., & Hafner, M. (2016). PAR-CLIP: A method for transcriptome-wide identification of RNA binding protein interaction sites. *Methods in Molecular Biology (Clifton, N.J.)*, *1358*, 153–173.

Dasti, A., Cid-Samper, F., Bechara, E., & Tartaglia, G. G. (2020). RNA-centric approaches to study RNA-protein interactions in vitro and in silico. *Methods* , *178*, 11–18.

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., Hilton, J. A., Jain, K., Baymuradov, U. K., Narayanan, A. K., Onate, K. C., Graham, K., Miyasato, S. R., Dreszer, T. R., Strattan, J. S., Jolanki, O., Tanaka, F. Y., & Cherry, J. M. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Research*, *46*(D1), D794–D801.

De, S., & Gorospe, M. (2017). Bioinformatic tools for analysis of CLIP ribonucleoprotein data. *Wiley Interdisciplinary Reviews. RNA*, *8*(4). https://doi.org/10.1002/wrna.1404

de Vasconcellos, J. F., Fasano, R. M., Lee, Y. T., Kaushal, M., Byrnes, C., Meier, E. R., Anderson, M., Rabel, A., Braylan, R., Stroncek, D. F., & Miller, J. L. (2014). LIN28A expression reduces sickling of cultured human erythrocytes. *PloS One*, *9*(9), e106924.

Ding, Y., Chan, C. Y., & Lawrence, C. E. (2004). Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Research*, *32*(Web Server issue), W135-41.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, *29*(1), 15–21.

Docherty, C. K., Salt, I. P., & Mercer, J. R. (2016). Lin28A induces energetic switching to glycolytic metabolism in human embryonic kidney cells. *Stem Cell Research & Therapy*, *7*(1), 78.

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., Yeo, G. W., Graveley, B. R., & Burge, C. B. (2018). Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Molecular Cell*, *70*(5), 854-867.e9.

Ellington, A. D., & Szostak, J. W. (1990). In vitro selection of RNA molecules that bind specific ligands. *Nature*, *346*(6287), 818–822.

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74.

Eshleman, N., Luo, X., Capaldi, A., & Buchan, J. R. (2020). Alterations of signaling pathways in response to chemical perturbations used to measure mRNA decay rates in yeast. *RNA* , *26*(1), 10–18.

Faas, L., Warrander, F. C., Maguire, R., Ramsbottom, S. A., Quinn, D., Genever, P., & Isaacs, H. V. (2013). Lin28 proteins are required for germ layer specification in Xenopus. *Development (Cambridge, England)*, *140*(5), 976–986.

Fathi, A., Eisa-Beygi, S., & Baharvand, H. (2017). Signaling Molecules Governing Pluripotency and Early Lineage Commitments in Human Pluripotent Stem Cells. *Cell Journal*, *19*(2), 194–203.

Feng, H., Bao, S., Rahman, M. A., Weyn-Vanhentenryck, S. M., Khan, A., Wong, J., Shah, A., Flynn, E. D., Krainer, A. R., & Zhang, C. (2019). Modeling RNA-Binding Protein Specificity In Vivo by Precisely Registering Protein-RNA Crosslink Sites. *Molecular Cell*, *74*(6), 1189-1204.e6.

Feng, H., Lu, X.-J., Liu, L., Ustianenko, D., & Zhang, C. (2022). Structure-based prediction and characterization of photo-crosslinking in native protein-RNA complexes. In *bioRxiv* (p. 2022.06.02.494568). https://doi.org/10.1101/2022.06.02.494568

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

Fu, M., & Blackshear, P. J. (2017). RNA-binding proteins in immune regulation: a focus on CCCH zinc finger proteins. *Nature Reviews. Immunology*, *17*(2), 130–143.

Galarneau, A., & Richard, S. (2005). Target RNA motif and target mRNAs of the Quaking STAR protein. *Nature Structural & Molecular Biology*, *12*(8), 691–698.

Garzia, A., Meyer, C., Morozov, P., Sajek, M., & Tuschl, T. (2017). Optimization of PAR-CLIP for transcriptome-wide identification of binding sites of RNA-binding proteins. *Methods (San Diego, Calif.)*, *118–119*, 24–40.

Gebauer, F., Schwarzl, T., Valcárcel, J., & Hentze, M. W. (2020). RNA-binding proteins in human genetic disease. *Nature Reviews. Genetics*. https://doi.org/10.1038/s41576-020-00302-y

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

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

Grabski, A. C. (2009). Advances in preparation of biological extracts for protein purification. *Methods in Enzymology*, *463*, 285–303.

Gueroussov, S., Weatheritt, R. J., O'Hanlon, D., Lin, Z.-Y., Narula, A., Gingras, A.-C., & Blencowe, B. J. (2017). Regulatory Expansion in Mammals of Multivalent hnRNP Assemblies that Globally Control Alternative Splicing. *Cell*, *170*(2), 324-339.e23.

Haberman, N., Huppertz, I., Attig, J., König, J., Wang, Z., Hauer, C., Hentze, M. W., Kulozik, A. E., Le Hir, H., Curk, T., Sibley, C. R., Zarnack, K., & Ule, J. (2017). Insights into the design and interpretation of iCLIP experiments. *Genome Biology*, *18*(1), 7.

Hafner, M., Katsantoni, M., Köster, T., Marks, J., Mukherjee, J., Staiger, D., Ule, J., & Zavolan, M. (2021). CLIP and complementary methods. *Nature Reviews Methods Primers*, *1*(1), 1–23.

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jr, Jungkamp, A.-C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M., & Tuschl, T. (2010). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, *141*(1), 129–141.

Hallegger, M., Chakrabarti, A. M., Lee, F. C. Y., Lee, B. L., Amalietti, A. G., Odeh, H. M., Copley, K. E., Rubien, J. D., Portz, B., Kuret, K., Huppertz, I., Rau, F., Patani, R., Fawzi, N. L., Shorter, J., Luscombe, N. M., & Ule, J. (2021). TDP-43 condensation properties specify its RNA-binding and regulatory repertoire. *Cell*. https://doi.org/10.1016/j.cell.2021.07.018

Hamilton, W. B., Mosesson, Y., Monteiro, R. S., Emdal, K. B., Knudsen, T. E., Francavilla, C., Barkai, N., Olsen, J. V., & Brickman, J. M. (2019). Dynamic lineage priming is driven via direct enhancer regulation by ERK. *Nature*, *575*(7782), 355–360.

Hanna, J., Saha, K., Pando, B., van Zon, J., Lengner, C. J., Creyghton, M. P., van Oudenaarden, A., & Jaenisch, R. (2009). Direct cell reprogramming is a stochastic process amenable to acceleration. *Nature*, *462*(7273), 595–601.

He, S., Valkov, E., Cheloufi, S., & Murn, J. (2023). The nexus between RNA-binding proteins and their effectors. *Nature Reviews. Genetics*, *24*(5), 276–294.

Heinemann, U., & Roske, Y. (2021). Cold-shock domains-abundance, structure, properties, and nucleic-acid binding. *Cancers*, *13*(2), 190.

Heo, I., Joo, C., Cho, J., Ha, M., Han, J., & Kim, V. N. (2008). Lin28 mediates the terminal uridylation of let-7 precursor MicroRNA. *Molecular Cell*, *32*(2), 276–284.

Heo, I., Joo, C., Kim, Y.-K., Ha, M., Yoon, M.-J., Cho, J., Yeom, K.-H., Han, J., & Kim, V. N. (2009). TUT4 in concert with Lin28 suppresses microRNA biogenesis through pre-microRNA uridylation. *Cell*, *138*(4), 696–708.

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

Hoell, J. I., Larsson, E., Runge, S., Nusbaum, J. D., Duggimpudi, S., Farazi, T. A., Hafner, M., Borkhardt, A., Sander, C., & Tuschl, T. (2011). RNA targets of wild-type and mutant FET family proteins. *Nature Structural & Molecular Biology*, *18*(12), 1428–1431.

Hofacker, I. L. (2003). Vienna RNA secondary structure server. *Nucleic Acids Research*, *31*(13), 3429–3431.

Hollmann, N. M., Jagtap, P. K. A., Linse, J.-B., Ullmann, P., Payr, M., Murciano, B., Simon, B., Hub, J. S., & Hennig, J. (2023). Upstream of N-Ras C-terminal cold shock domains mediate poly(A) specificity in a novel RNA recognition mode and bind poly(A) binding protein. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkac1277

Horlacher, M., Wagner, N., Moyon, L., Kuret, K., Goedert, N., Salvatore, M., Ule, J., Gagneur, J., Winther, O., & Marsico, A. (2023). Towards in silico CLIP-seq: predicting protein-RNA interaction via sequence-to-signal learning. *Genome Biology*, *24*(1), 180.

Hu, Z., Ma, J., Yue, H., Luo, Y., Li, X., Wang, C., Wang, L., Sun, B., Chen, Z., Wang, L., & Gu, Y. (2022). Involvement of LIN28A in Wnt-dependent regulation of hippocampal neurogenesis in the aging brain. *Stem Cell Reports*. https://doi.org/10.1016/j.stemcr.2022.05.016

Huppertz, I., Attig, J., D'Ambrogio, A., Easton, L. E., Sibley, C. R., Sugimoto, Y., Tajnik, M., König, J., & Ule, J. (2014). iCLIP: protein-RNA interactions at nucleotide resolution. *Methods* , *65*(3), 274–287.

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

Jolma, A., Zhang, J., Mondragón, E., Morgunova, E., Kivioja, T., Laverty, K. U., Yin, Y., Zhu, F., Bourenkov, G., Morris, Q., Hughes, T. R., Maher, L. J., 3rd, & Taipale, J. (2020). Binding specificities of human RNA-binding proteins toward structured and linear RNA sequences. *Genome Research*, *30*(7), 962–973.

Katsantoni, M., van Nimwegen, E., & Zavolan, M. (2023). Improved analysis of (e)CLIP data with RCRUNCH yields a compendium of RNA-binding protein binding sites and motifs. *Genome Biology*, *24*(1), 77.

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

Knight, R. (n.d.). *scikit-bio*. Github. Retrieved August 17, 2023, from https://github.com/biocore/scikit-bio

Knörlein, A., Sarnowski, C. P., de Vries, T., Stoltz, M., Götze, M., Aebersold, R., Allain, F. H.-T., Leitner, A., & Hall, J. (2022). Nucleotide-amino acid π-stacking interactions initiate photo cross-linking in RNA-protein complexes. *Nature Communications*, *13*(1), 2719.

König, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M., & Ule, J. (2010). iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nature Structural & Molecular Biology*, *17*(7), 909–915.

Konig, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M., & Ule, J. (2011). iCLIP--transcriptome-wide mapping of protein-RNA interactions with individual nucleotide resolution. *Journal of Visualized Experiments: JoVE*, *50*. https://doi.org/10.3791/2638

Krakau, S., Richard, H., & Marsico, A. (2017). PureCLIP: capturing target-specific protein-RNA interaction footprints from single-nucleotide CLIP-seq data. *Genome Biology*, *18*(1), 240.

Kuret, K. (2023). *ulelab/cluster_kmers: v0.0.0*. Zenodo. https://doi.org/10.5281/ZENODO.8386584

Kuret, K., Amalietti, A. G., Jones, D. M., Capitanchik, C., & Ule, J. (2022). Positional motif analysis reveals the extent of specificity of protein-RNA interactions observed by CLIP. *Genome Biology*, *23*(1), 1–34.

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

Langmead, B., Trapnell, C., Pop, M., & Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, *10*(3), R25.

Laverty, K. U., Jolma, A., Pour, S. E., Zheng, H., Ray, D., Morris, Q., & Hughes, T. R. (2022). PRIESSTESS: interpretable, high-performing models of the sequence and structure preferences of RNA-binding proteins. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkac694

Lee, A. S. Y., Kranzusch, P. J., & Cate, J. H. D. (2015). eIF3 targets cell-proliferation messenger RNAs for translational activation or repression. *Nature*, *522*(7554), 111–114.

Lee, F. C. Y., Chakrabarti, A. M., Hänel, H., Monzón-Casanova, E., Hallegger, M., Militti, C., Capraro, F., Sadée, C., Toolan-Kerr, P., Wilkins, O., Turner, M., König, J., Sibley, C. R., & Ule, J. (2021). An improved iCLIP protocol. In *bioRxiv* (p. 2021.08.27.457890). https://doi.org/10.1101/2021.08.27.457890

Lee, F. C. Y., & Ule, J. (2018). Advances in CLIP Technologies for Studies of Protein-RNA Interactions. *Molecular Cell*, *69*(3), 354–369.

Leibovich, L., Paz, I., Yakhini, Z., & Mandel-Gutfreund, Y. (2013). DRIMust: a web server for discovering rank imbalanced motifs using suffix trees. *Nucleic Acids Research*, *41*(Web Server issue), W174-9.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & 1000 Genome Project Data Processing Subgroup. (2009). The

Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, *25*(16), 2078–2079.

Li, M., & Izpisua Belmonte, J. C. (2018). Deconstructing the pluripotency gene regulatory network. *Nature Cell Biology*, *20*(4), 382–392.

Li, X., Quon, G., Lipshitz, H. D., & Morris, Q. (2010). Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. *RNA (New York, N.Y.)*, *16*(6), 1096–1107.

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., Darnell, J. C., & Darnell, R. B. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, *456*(7221), 464–469.

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

López de Silanes, I., Galbán, S., Martindale, J. L., Yang, X., Mazan-Mamczarz, K., Indig, F. E., Falco, G., Zhan, M., & Gorospe, M. (2005). Identification and functional outcome of mRNAs associated with RNA-binding protein TIA-1. *Molecular and Cellular Biology*, *25*(21), 9520–9531.

Lorenz, R., Bernhart, S. H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology: AMB*, *6*, 26.

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550.

Ma, X., Li, C., Sun, L., Huang, D., Li, T., He, X., Wu, G., Yang, Z., Zhong, X., Song, L., Gao, P., & Zhang, H. (2014). Lin28/let-7 axis regulates aerobic glycolysis and cancer progression via PDK1. *Nature Communications*, *5*(1), 5212.

Maaskola, J., & Rajewsky, N. (2014). Binding site discovery from nucleic acid sequences by discriminative learning of hidden Markov models. *Nucleic Acids Research*, *42*(21), 12995–13011.

Marinus, T., Fessler, A. B., Ogle, C. A., & Incarnato, D. (2021). A novel SHAPE reagent enables the analysis of RNA structure in living cells with unprecedented accuracy. *Nucleic Acids Research*, *49*(6), e34.

Maris, C., Dominguez, C., & Allain, F. H.-T. (2005). The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *The FEBS Journal*, *272*(9), 2118–2131.

Martin, G., Gruber, A. R., Keller, W., & Zavolan, M. (2012). Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length. *Cell Reports*, *1*(6), 753–763.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.Journal*, *17*(1), 10.

Maticzka, D., Lange, S. J., Costa, F., & Backofen, R. (2014). GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biology*, *15*(1), R17.

Modic, M., de Los Mozos, I. R., Steinhauser, S., van Genderen, E., Schirge, S., Bergant, V., Ryan, J., Mulholland, C. B., Faraway, R., Lee, F. C. Y., Klobučar, T., Merl-Pham, J., Hauck, S. M., Drukker, M., Bultmann, S., Leonhardt, H., Lickert, H., Luscombe, N. M., ten Berge, D., & Ule, J. (2021). Epiblast morphogenesis is controlled by selective mRNA decay triggered by LIN28A relocation. In *Cold Spring Harbor Laboratory* (p. 2021.03.15.433780). https://doi.org/10.1101/2021.03.15.433780

Modic, M., Kuret, K., Steinhauser, S., Faraway, R., van Genderen, E., Lee, F., Mozos, I. R. de L., Vičič, Ž., Novljan, J., Berge, D. T., Luscombe, N., & Ule, J. (2023). Poised PABP-RNA hubs implement signal-dependent mRNA decay in development. In *Research Square*. https://doi.org/10.21203/rs.3.rs-3227673/v1

Morandi, E., van Hemert, M. J., & Incarnato, D. (2022). SHAPE-guided RNA structure homology search and motif discovery. *Nature Communications*, *13*(1), 1722.

Mukherjee, N., Wessels, H.-H., Lebedeva, S., Sajek, M., Ghanbari, M., Garzia, A., Munteanu, A., Yusuf, D., Farazi, T., Hoell, J. I., Akat, K. M., Akalin, A., Tuschl, T., & Ohler, U. (2019). Deciphering human ribonucleoprotein regulatory networks. *Nucleic Acids Research*, *47*(2), 570–581.

Nabeel-Shah, S., & Greenblatt, J. F. (2023). Revised iCLIP-seq protocol for profiling RNA-protein interaction sites at individual nucleotide resolution in living cells. *Bio-Protocol*, *13*(11), e4688.

Nam, Y., Chen, C., Gregory, R. I., Chou, J. J., & Sliz, P. (2011). Molecular basis for interaction of let-7 microRNAs with Lin28. *Cell*, *147*(5), 1080–1091.

Neagu, A., van Genderen, E., Escudero, I., Verwegen, L., Kurek, D., Lehmann, J., Stel, J., Dirks, R. A. M., van Mierlo, G., Maas, A., Eleveld, C., Ge, Y., den Dekker, A. T., Brouwer, R. W. W., van IJcken, W. F. J., Modic, M., Drukker, M., Jansen, J. H., Rivron, N. C., … Ten Berge, D. (2020). In vitro capture and characterization of embryonic rosette-stage pluripotency between naive and primed states. *Nature Cell Biology*, *22*(5), 534–545.

Nicastro, G., Taylor, I. A., & Ramos, A. (2015). KH-RNA interactions: back in the groove. *Current Opinion in Structural Biology*, *30*, 63–70.

Orenstein, Y., Wang, Y., & Berger, B. (2016). RCK: accurate and efficient inference of sequence- and structure-based protein-RNA binding models from RNAcompete data. *Bioinformatics (Oxford, England)*, *32*(12), i351–i359.

Pan, X., & 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), 136.

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research: JMLR*.

Peng, S., Chen, L.-L., Lei, X.-X., Yang, L., Lin, H., Carmichael, G. G., & Huang, Y. (2011). Genome-wide studies reveal that Lin28 enhances the translation of genes important for growth and survival of human embryonic stem cells. *Stem Cells (Dayton, Ohio)*, *29*(3), 496–504.

Perez-Perri, J. I., Rogell, B., Schwarzl, T., Stein, F., Zhou, Y., Rettel, M., Brosig, A., & Hentze, M. W. (2018). Discovery of RNA-binding proteins and characterization of their dynamic responses by enhanced RNA interactome capture. *Nature Communications*, *9*(1), 4408.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Meng, E. C., Couch, G. S., Croll, T. I., Morris, J. H., & Ferrin, T. E. (2021). UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Science: A Publication of the Protein Society*, *30*(1), 70–82.

Pietrosanto, M., Ausiello, G., & Helmer-Citterich, M. (2021). Motif Discovery from CLIP Experiments. *Methods in Molecular Biology* , *2284*, 43–50.

Piskounova, E., Polytarchou, C., Thornton, J. E., LaPierre, R. J., Pothoulakis, C., Hagan, J. P., Iliopoulos, D., & Gregory, R. I. (2011). Lin28A and Lin28B inhibit let-7 microRNA biogenesis by distinct mechanisms. *Cell*, *147*(5), 1066–1079.

Polesskaya, A., Cuvellier, S., Naguibneva, I., Duquet, A., Moss, E. G., & Harel-Bellan, A. (2007). Lin-28 binds IGF-2 mRNA and participates in skeletal myogenesis by increasing translation efficiency. *Genes & Development*, *21*(9), 1125–1138.

Qi, Y., Wang, M., & Jiang, Q. (2022). PABPC1--mRNA stability, protein translation and tumorigenesis. *Frontiers in Oncology*, *12*, 1025291.

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* , *26*(6), 841–842.

Rastogi, C., Rube, H. T., Kribelbauer, J. F., Crocker, J., Loker, R. E., Martini, G. D., Laptenko, O., Freed-Pastor, W. A., Prives, C., Stern, D. L., Mann, R. S., & Bussemaker, H. J. (2018). Accurate and sensitive quantification of protein-DNA binding affinity. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(16), E3692–E3701.

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

Ray, D., Kazan, H., Cook, K. B., Weirauch, M. T., Najafabadi, H. S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., Na, H., Irimia, M., Matzat, L. H., Dale, R. K., Smith, S. A., Yarosh, C. A., Kelly, S. M., Nabet, B., Mecenas, D., … Hughes, T. R. (2013). A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, *499*(7457), 172–177.

Ray, D., Laverty, K. U., Jolma, A., Nie, K., Samson, R., Pour, S. E., Tam, C. L., von Krosigk, N., Nabeel-Shah, S., Albu, M., Zheng, H., Perron, G., Lee, H., Najafabadi, H., Blencowe, B., Greenblatt, J., Morris, Q., & Hughes, T. R. (2023). RNA-binding proteins that lack canonical RNA-binding domains are rarely sequence-specific. *Scientific Reports*, *13*(1), 5238.

Respuela, P., Nikolić, M., Tan, M., Frommolt, P., Zhao, Y., Wysocka, J., & Rada-Iglesias, A. (2016). Foxd3 promotes exit from naive pluripotency through enhancer decommissioning and inhibits germline specification. *Cell Stem Cell*, *18*(1), 118–133.

Romo, L., Findlay, S. D., & Burge, C. B. (2023). Regulatory features aid interpretation of 3'UTR Variants. In *bioRxiv* (p. 2023.08.01.551549). https://doi.org/10.1101/2023.08.01.551549

Rossbach, O., Hung, L.-H., Khrameeva, E., Schreiner, S., König, J., Curk, T., Zupan, B., Ule, J., Gelfand, M. S., & Bindereif, A. (2014). Crosslinking-immunoprecipitation (iCLIP) analysis reveals global regulatory roles of hnRNP L. *RNA Biology*, *11*(2), 146–155.

Roush, S., & Slack, F. J. (2008). The let-7 family of microRNAs. *Trends in Cell Biology*, *18*(10), 505–516.

Saito, M., Hess, D., Eglinger, J., Fritsch, A. W., Kreysing, M., Weinert, B. T., Choudhary, C., & Matthias, P. (2019). Acetylation of intrinsically disordered regions regulates phase separation. *Nature Chemical Biology*, *15*(1), 51–61.

Sasse, A., Laverty, K. U., Hughes, T. R., & Morris, Q. D. (2018). Motif models for RNA-binding proteins. *Current Opinion in Structural Biology*, *53*, 115–123.

Schueler, M., Munschauer, M., Gregersen, L. H., Finzel, A., Loewer, A., Chen, W., Landthaler, M., & Dieterich, C. (2014). Differential protein occupancy profiling of the mRNA transcriptome. *Genome Biology*, *15*(1), R15.

Schwarzl, T., Sahadevan, S., Lang, B., Miladi, M., Backofen, R., Huber, W., Hentze, M. W., & Tartaglia, G. G. (2022). Improved discovery of RNA-binding protein binding sites in eCLIP data using DEWSeq. In *bioRxiv* (p. 2022.11.15.516416). https://doi.org/10.1101/2022.11.15.516416

Shah, A., Qian, Y., Weyn-Vanhentenryck, S. M., & Zhang, C. (2017). CLIP Tool Kit (CTK): a flexible and robust pipeline to analyze CLIP sequencing data. *Bioinformatics* , *33*(4), 566–567.

Shah, K., He, S., Turner, D. J., Corbo, J., Rebbani, K., Bateman, J. M., Cheloufi, S., Igreja, C., Valkov, E., & Murn, J. (2023). A paradigm for regulation at the effector interface with RNA-binding proteins. In *bioRxiv*. https://doi.org/10.1101/2023.09.20.558714

Shyh-Chang, N., & Daley, G. Q. (2013). Lin28: primal regulator of growth and metabolism in stem cells. *Cell Stem Cell*, *12*(4), 395–406.

Shyh-Chang, N., Zhu, H., Yvanka de Soysa, T., Shinoda, G., Seligson, M. T., Tsanov, K. M., Nguyen, L., Asara, J. M., Cantley, L. C., & Daley, G. Q. (2013). Lin28 enhances tissue repair by reprogramming cellular metabolism. *Cell*, *155*(4), 778–792.

Smith, T., Heger, A., & Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Research*, *27*(3), 491–499.

Spassov, D. S., & Jurecic, R. (2003). The PUF family of RNA-binding proteins: does evolutionarily conserved structure equal conserved function? *IUBMB Life*, *55*(7), 359–366.

Stormo, G. D., Schneider, T. D., Gold, L., & Ehrenfeucht, A. (1982). Use of the 'Perceptron' algorithm to distinguish translational initiation sites inE. coli. *Nucleic Acids Research*, *10*(9), 2997–3011.

Sun, L., Fazal, F. M., Li, P., Broughton, J. P., Lee, B., Tang, L., Huang, W., Kool, E. T., Chang, H. Y., & Zhang, Q. C. (2019). RNA structure maps across mammalian cellular compartments. *Nature Structural & Molecular Biology*, *26*(4), 322–330.

Tareen, A., & Kinney, J. B. (2019). Logomaker: Beautiful sequence logos in python. In *bioRxiv* (p. 635029). https://doi.org/10.1101/635029

TextDistance. (2021 1). https://github.com/life4/textdistance/releases/tag/v.4.2.0

Tollervey, J. R., Curk, T., Rogelj, B., Briese, M., Cereda, M., Kayikci, M., König, J., Hortobágyi, T., Nishimura, A. L., Zupunski, V., Patani, R., Chandran, S., Rot, G., Zupan, B., Shaw, C. E., & Ule, J. (2011). Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nature Neuroscience*, *14*(4), 452–458.

Tsanov, K. M., Pearson, D. S., Wu, Z., Han, A., Triboulet, R., Seligson, M. T., Powers, J. T., Osborne, J. K., Kane, S., Gygi, S. P., Gregory, R. I., & Daley, G. Q. (2017). LIN28 phosphorylation by MAPK/ERK couples signalling to the post-transcriptional control of pluripotency. *Nature Cell Biology*, *19*(1), 60–67.

Tsialikas, J., & Romer-Seibert, J. (2015). LIN28: roles and regulation in development and beyond. *Development* , *142*(14), 2397–2404.

Tuerk, C., & Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, *249*(4968), 505–510.

Ule, J., Jensen, K. B., Ruggiu, M., Mele, A., Ule, A., & Darnell, R. B. (2003). CLIP identifies Nova-regulated RNA networks in the brain. *Science*, *302*(5648), 1212–1215.

Ule, J., Jensen, K., Mele, A., & Darnell, R. B. (2005). CLIP: a method for identifying protein-RNA interaction sites in living cells. *Methods (San Diego, Calif.)*, *37*(4), 376–386.

Uren, P. J., Bahrami-Samani, E., Burns, S. C., Qiao, M., Karginov, F. V., Hodges, E., Hannon, G. J., Sanford, J. R., Penalva, L. O. F., & Smith, A. D. (2012). Site identification in high-throughput RNA-protein interaction data. *Bioinformatics* , *28*(23), 3013–3020.

Ustianenko, D., Chiu, H.-S., Treiber, T., Weyn-Vanhentenryck, S. M., Treiber, N., Meister, G., Sumazin, P., & Zhang, C. (2018). LIN28 Selectively Modulates a Subclass of Let-7 MicroRNAs. *Molecular Cell*, *71*(2), 271-283.e5.

Van Nostrand, E. L., Freese, P., Pratt, G. A., Wang, X., Wei, X., Xiao, R., Blue, S. M., Chen, J.-Y., Cody, N. A. L., Dominguez, D., Olson, S., Sundararaman, B., Zhan, L., Bazile, C., Bouvrette, L. P. B., Bergalet, J., Duff, M. O., Garcia, K. E., Gelboin-Burkhart, C., … Yeo, G. W. (2020). A large-scale binding and functional map of human RNA-binding proteins. *Nature*, *583*(7818), 711–719.

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., Stanton, R., Rigo, F., Guttman, M., & Yeo, G. W. (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature Methods*, *13*(6), 508–514.

Vega-Sendino, M., Olbrich, T., Tillo, D., Tran, A. D., Domingo, C. N., Franco, M., FitzGerald, P. C., Kruhlak, M. J., & Ruiz, S. (2021). The ETS transcription factor ERF controls the exit from the naïve pluripotent state in a MAPK-dependent manner. *Science Advances*, *7*(40), eabg8306.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272.

Wang, H., Zhao, Q., Deng, K., Guo, X., & Xia, J. (2016). Lin28: an emerging important oncogene connecting several aspects of cancer. *Tumour Biology: The Journal of the International Society for Oncodevelopmental Biology and Medicine*, *37*(3), 2841–2848.

Wang, X., Yu, S., Lou, E., Tan, Y.-L., & Tan, Z.-J. (2023). RNA 3D Structure Prediction: Progress and Perspective. *Molecules* , *28*(14). https://doi.org/10.3390/molecules28145532

Wang, Z., Kayikci, M., Briese, M., Zarnack, K., Luscombe, N. M., Rot, G., Zupan, B., Curk, T., & Ule, J. (2010). iCLIP predicts the dual splicing effects of TIA-RNA interactions. *PLoS Biology*, *8*(10), e1000530.

Webster, M. W., Chen, Y.-H., Stowell, J. A. W., Alhusaini, N., Sweet, T., Graveley, B. R., Coller, J., & Passmore, L. A. (2018). MRNA deadenylation is coupled to translation rates by the differential activities of Ccr4-not nucleases. *Molecular Cell*, *70*(6), 1089-1100.e8.

West, C., Chakrabarti, N., Patel, H., Ewels, P., & Capitanchik, C. (2021). *nf-core/clipseq: nf-core/clipseq 1.0.0 - Ianthine Pelican* (1.0.0) [Computer software]. https://doi.org/10.5281/zenodo.4723017

Wilbert, M. L., Huelga, S. C., Kapeli, K., Stark, T. J., Liang, T. Y., Chen, S. X., Yan, B. Y., Nathanson, J. L., Hutt, K. R., Lovci, M. T., Kazan, H., Vu, A. Q., Massirer, K. B., Morris, Q., Hoon, S., & Yeo, G. W. (2012). LIN28 binds messenger RNAs at GGAGA motifs and regulates splicing factor abundance. *Molecular Cell*, *48*(2), 195–206.

Wilkins, O. (2021). *ulelab/ultraplex: Ultraplex release.* https://doi.org/10.5281/zenodo.4651285

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018.

Wolin, E., Guo, J. K., Blanco, M. R., Perez, A. A., Goronzy, I. N., Abdou, A. A., Gorhe, D., Guttman, M., & Jovanovic, M. (2023). SPIDR: a highly multiplexed method for mapping RNA-protein interactions uncovers a potential mechanism for selective translational suppression upon cellular stress. In *bioRxiv* (p. 2023.06.05.543769). https://doi.org/10.1101/2023.06.05.543769

Wu, K., Ahmad, T., & Eri, R. (2022). LIN28A: A multifunctional versatile molecule with future therapeutic potential. *World Journal of Biological Chemistry*, *13*(2), 35–46.

Yamamoto, H., Uchida, Y., Kurimoto, R., Chiba, T., Matsushima, T., Ito, Y., Inotsume, M., Miyata, K., Watanabe, K., Inada, M., Goshima, N., Uchida, T., & Asahara, H. (2022). RNA-binding protein LIN28A upregulates transcription factor HIF1 α by post-transcriptional regulation via direct binding to UGAU motifs. *The Journal of Biological Chemistry*, 102791.

Yang, D. H., & Moss, E. G. (2003). Temporally regulated expression of Lin-28 in diverse tissues of the developing mouse. *Gene Expression Patterns : GEP*, *3*(6), 719–726.

Yang, P., Humphrey, S. J., Cinghu, S., Pathania, R., Oldfield, A. J., Kumar, D., Perera, D., Yang, J. Y. H., James, D. E., Mann, M., & Jothi, R. (2019). Multi-omic Profiling Reveals Dynamics of the Phased Progression of Pluripotency. *Cell Systems*, *8*(5), 427-445.e10.

Yang, R., Gaidamakov, S. A., Xie, J., Lee, J., Martino, L., Kozlov, G., Crawford, A. K., Russo, A. N., Conte, M. R., Gehring, K., & Maraia, R. J. (2011). La-related protein 4 binds poly(A), interacts with the poly(A)-binding protein MLLE domain via a variant PAM2w motif, and can promote mRNA stability. *Molecular and Cellular Biology*, *31*(3), 542–556.

Yeo, G. W., Coufal, N. G., Liang, T. Y., Peng, G. E., Fu, X.-D., & Gage, F. H. (2009). An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells. *Nature Structural & Molecular Biology*, *16*(2), 130–137.

Yi, H., Park, J., Ha, M., Lim, J., Chang, H., & Kim, V. N. (2018). PABP cooperates with the CCR4-NOT complex to promote mRNA deadenylation and block precocious decay. *Molecular Cell*, *70*(6), 1081-1088.e5.

Yu, J., Vodyanik, M. A., Smuga-Otto, K., Antosiewicz-Bourget, J., Frane, J. L., Tian, S., Nie, J., Jonsdottir, G. A., Ruotti, V., Stewart, R., Slukvin, I. I., & Thomson, J. A. (2007). Induced pluripotent stem cell lines derived from human somatic cells. *Science (New York, N.Y.)*, *318*(5858), 1917–1920.

Yu, N.-K., McClatchy, D. B., Diedrich, J. K., Romero, S., Choi, J.-H., Martínez-Bartolomé, S., Delahunty, C. M., Muotri, A. R., & Yates, J. R., 3rd. (2021). Interactome analysis illustrates diverse gene regulatory processes associated with LIN28A in human iPS cell-derived neural progenitor cells. *IScience*, *24*(11), 103321.

Zampetaki, A., Albrecht, A., & Steinhofel, K. (2018). Long non-coding RNA structure and function: Is there a link? *Frontiers in Physiology*, *9*. https://doi.org/10.3389/fphys.2018.01201

Zarnack, K., König, J., Tajnik, M., Martincorena, I., Eustermann, S., Stévant, I., Reyes, A., Anders, S., Luscombe, N. M., & Ule, J. (2013). Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell*, *152*(3), 453–466.

Zarnegar, B. J., Flynn, R. A., Shen, Y., Do, B. T., Chang, H. Y., & Khavari, P. A. (2016). irCLIP platform for efficient characterization of protein-RNA interactions. *Nature Methods*, *13*(6), 489–492.

Zhang, J., Ratanasirintrawoot, S., Chandrasekaran, S., Wu, Z., Ficarro, S. B., Yu, C., Ross, C. A., Cacchiarelli, D., Xia, Q., Seligson, M., Shinoda, G., Xie, W., Cahan, P., Wang, L., Ng, S.-C., Tintara, S., Trapnell, C., Onder, T., Loh, Y.-H., … Daley, G. Q. (2016). LIN28 Regulates Stem Cell Metabolism and Conversion to Primed Pluripotency. *Cell Stem Cell*, *19*(1), 66–80.

Zhao, W., Zhang, S., Zhu, Y., Xi, X., Bao, P., Ma, Z., Kapral, T. H., Chen, S., Zagrovic, B., Yang, Y. T., & Lu, Z. J. (2022). POSTAR3: an updated platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins. *Nucleic Acids Research*, *50*(D1), D287–D294.

Zhu, A., Ibrahim, J. G., & Love, M. I. (2019). Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences. *Bioinformatics (Oxford, England)*, *35*(12), 2084–2092.

Zhu, H., Shah, S., Shyh-Chang, N., Shinoda, G., Einhorn, W. S., Viswanathan, S. R., Takeuchi, A., Grasemann, C., Rinn, J. L., Lopez, M. F., Hirschhorn, J. N., Palmert, M. R., & Daley, G. Q. (2010). Lin28a transgenic mice manifest size and puberty phenotypes identified in human genetic association studies. *Nature Genetics*, *42*(7), 626–630.

Zou, H., Luo, J., Guo, Y., Liu, Y., Wang, Y., Deng, L., & Li, P. (2022). RNA-binding protein complex LIN28/MSI2 enhances cancer stem cell-like properties by modulating Hippo-YAP1 signaling and independently of Let-7. *Oncogene*, *41*(11), 1657–1672.

Bibliography

Publications Related to the Thesis

Journal Articles

Kuret, K., Amalietti, A. G., Jones, D. M., Capitanchik, C., & Ule, J. (2022). Positional motif analysis reveals the extent of specificity of protein-RNA interactions observed by CLIP. Genome Biology, 23(1), 1–34.

The following paper was accepted at Nature Structure and Molecular Biology and is expected to be published in January 2024. Until the publication, the work is available as a pre-print:

Modic, M., Kuret, K., Steinhauser, S., Faraway, R., van Genderen, E., Lee, F., Mozos, I. R. de L., Vičič, Ž., Novljan, J., Berge, D. T., Luscombe, N., & Ule, J. (2023). Poised PABP-RNA hubs implement signal-dependent mRNA decay in development. In *Research Square*. https://doi.org/10.21203/rs.3.rs-3227673/v1

Other Publications

Journal Articles

Hallegger, M., Chakrabarti, A. M., Lee, F. C. Y., Lee, B. L., Amalietti, A. G., Odeh, H. M., Copley, K. E., Rubien, J. D., Portz, B., Kuret, K., Huppertz, I., Rau, F., Patani, R., Fawzi, N. L., Shorter, J., Luscombe, N. M., & Ule, J. (2021). TDP-43 condensation properties specify its RNA-binding and regulatory repertoire. Cell. https://doi.org/10.1016/j.cell.2021.07.018

Horlacher, M., Wagner, N., Moyon, L., Kuret, K., Goedert, N., Salvatore, M., Ule, J., Gagneur, J., Winther, O., & Marsico, A. (2023). Towards in silico CLIP-seq: predicting protein-RNA interaction via sequence-to-signal learning. Genome Biology, 24(1), 180.

Preprints

Capitanchik, C., Ireland, S., Harston, A., Cheshire, C., Marc Jones, D., Lee, F. C. Y., de los Mozos, I. R., Iosub, I. A., Kuret, K., Faraway, R., Wilkins, O. G., Arora, R., Hallegger, M., Modic, M., Chakrabarti, A. M., Luscombe, N. M., & Ule, J. (2023). Flow: a web platform and open database to analyse, store, curate and share bioinformatics data at scale. In bioRxiv (p. 2023.08.22.544179). https://doi.org/10.1101/2023.08.22.544179

Biography

Klara Kuret was born in Ljubljana on the 5th of September 1995. In 2014, she enrolled in the Biochemistry Bachelor's programme at the Faculty of Chemistry and Chemical Technology, University of Ljubljana. She completed her BSc studies in 2017 by defending her thesis titled "Preparation of recombinant β-catenin and FHL2 and optimisation of the production process", which she prepared under the supervision of Prof. Brigita Lenarčič and Dr. Aljaž Gaber.

Between 2017 and 2019, she continued her studies at The Faculty of Pharmacy, University of Ljubljana, earning a Master's degree in Industrial Pharmacy on the 17th of June 2020. She conducted research for her Master's thesis abroad—in the Laboratory for Structural Biology at Åbo Akademi University, in Turku, Finland. Under the supervision of her mentors Prof. Outi Salo-Ahen and Assistant Prof. Marko Jukić, she prepared her thesis titled "*In silico* search for novel bacterial RNA polymerase inhibitors", for which she employed high-throughput computational screening approaches and molecular simulations. During that period, she was introduced to the scientific fields of computational biology and bioinformatics, in which she later continued her work.

In 2019, she joined the Laboratory for RNA Networks at the National Institute of Chemistry in Ljubljana, which was established that same year by Prof. Jernej Ule. She began developing and applying bioinformatic approaches to study cellular interactions between proteins and RNA, specialising in the analysis of CLIP data. In 2020, she continued her work in the same lab as a PhD student under the supervision of Prof. Jernej Ule and Dr. Miha Modic. She enrolled in the doctoral study programme at the Jožef Stefan International Postgraduate School. During her doctoral studies, she produced the body of research presented in this thesis; as well as other papers published in collaboration with colleagues at the Francis Crick Institute and at the Helmholtz Zentrum München. In 2023, she was awarded a national scholarship by the L'Oreal-Unesco "For Women In Science" programme for her doctoral research. In addition to her publications, she also presented her work at workshops and international conferences.