PhD Program in Biomolecular Sciences

Department of Cellular, Computational and Integrative Biology – CIBIO

37th Cycle

Computational analysis of cooperation and competition mechanisms between RNA-binding proteins

Tutor/Advisor

Prof. Erik Dassi

University of Trento

Ph.D. Thesis of

Carlos Francisco Peula Cordero

University of Trento

Academic Year 2021-2022

DECLARATION OF ORIGINAL AUTHORSHIP:

I, Carlos Francisco Peula Cordero, confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Trento, 20/12/2024 Carlos Francisco Peula Cordero

CONTENT INDEX

TABLE OF FIGURES

Abstract

Gene expression is regulated at five different levels, namely epigenetic, transcriptional, post-transcriptional, translational and post-translational. The post-transcriptional regulatory layer accounts for all the events taking place between transcription and translation and involves a myriad of trans and cis factors densely interconnected. Among the regulatory trans-elements forming this network, RNA-binding proteins (RBPs) play key roles in post-transcriptional regulation of gene expression (PTR). Understanding RBPs behavior is therefore crucial given their participation in all stages of the mRNA life cycle. Although an increasing amount of work has highlighted the relevance of RBP-RBP interactions for PTR, only recent studies have focused on the development of comprehensive approaches for prediction of potential RBP-RBP interactions previously unreported. Despite these steps toward a comprehensive RBP-RBP interaction map, the underlying driver mechanisms of such interactions are still not well characterized. The present work aims thus at providing insights into the complexity of this highly interconnected regulatory network. Leveraging

positional information of RNA-binding sites from crosslinking and immunoprecipitation assays, we provide an initial approach to discern between RBP cooperation and competition. In addition, we evaluated the impact of multiple features on the RBP-RBP interaction modalities, eventually directing our focus towards the role of post-translational modifications (PTMs) in these mechanisms. Our findings suggest that potential PTM sites influence the modality of RBP-RBP interactions thus being a candidate determinant factor of cooperative and competitive behaviors.

Introduction

1. Breaking down the sources of cell variability.

The human body of an average adult consists of approximately 30 trillion cells. Although the genetic information contained in DNA is the same for all somatic cells, our soma is composed of more than 200 different cell types (Sender and Milo 2021; Bianconi et al. 2013; Mazzarello 1999). This cell functional diversity is mostly accomplished through the differential expression of the common pool of genes, which, in turn, allows each cell type to acquire its distinctive properties. More precisely, cell type-specific gene expression is achieved through a number of mechanisms, at different levels, that do not alter the primary sequence of DNA.

Although much of this variability can be explained by epigenetic and transcriptional regulatory mechanisms, there is a vast number of post-transcriptional events which are also determinant for differential gene expression (Corbett 2018). Moreover, the translation of messenger RNAs (mRNAs) to proteins is also subjected to regulation by multiple factors that can promote or reduce the ribosomal translation efficiency. Lastly, after translation takes place, many proteins are potential substrates for enzymes that add and remove functional groups to their targets. These act as modulators of their structure and functionality, and represent an additional regulatory layer at the post-translational level (Velázquez-Cruz et al. 2021).

To sum up, all the regulatory layers of gene expression together contribute to the differential expression of genes, consequently generating phenotypic variability among somatic cells (**Figure 1**). Therefore, this introduction is meant to provide an overview of the main aspects of each layer, with particular focus on the post-transcriptional and post-translational levels.

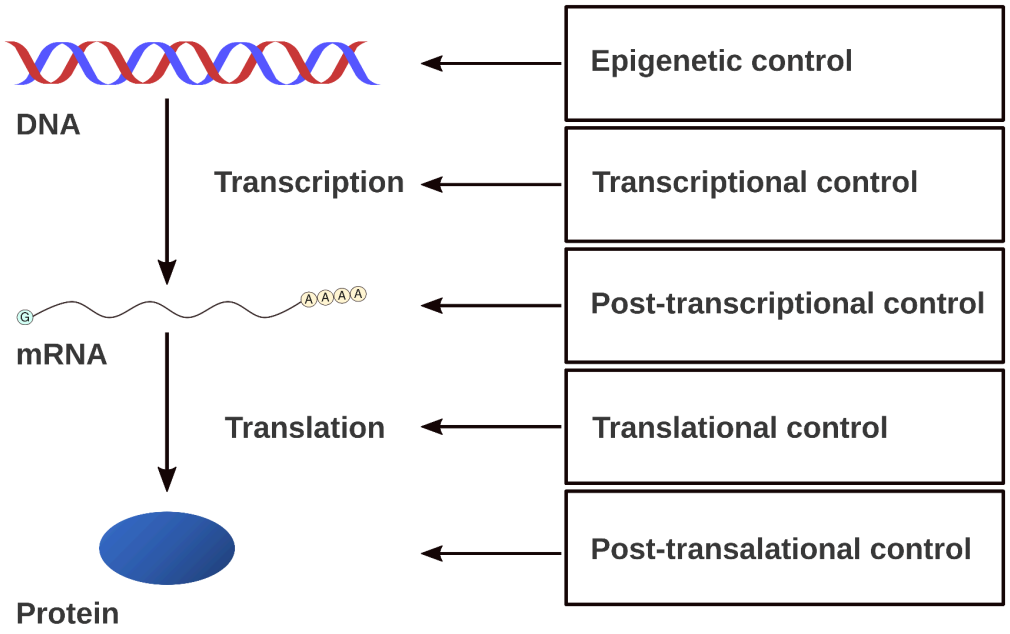


Figure 1: Schematic representation of the central dogma of molecular biology. *The diagram shows the five different regulatory layers controlling the gene expression in eukaryotic cells.*

2. A synopsis of the transcriptional modulation of gene expression

The transcriptional regulatory layer encompasses a series of mechanisms taking place within the nucleus which directly or indirectly control the transcription rate of genes. In eukaryotic cells, the RNA polymerase II (RNA Pol II) transcribes protein-coding genes into mRNAs. In addition, RNA Pol II also synthesizes long non coding RNA (lncRNA) micro RNA (miRNA) and small nuclear RNAs (snRNAs). In this context, the transcriptional control occurs through the action of transcription factors (TFs) which act as regulators of transcription by binding to the promoter regions of genes, thus facilitating or preventing the RNA Pol II docking (Archuleta, Goodrich, and Kugel 2024).

While the RNA Pol II can synthesize RNA using DNA as a template, it lacks sequence-specific DNA binding capacity. Therefore, it requires general TFs as a guide for binding at the promoter region of genes containing the transcriptional start site. The general TFs (TFIIA, TFIIB, TFIID, TFIIE, TFIIF, and TFIIH) assemble with the RNA Pol II into a preinitiation complex (Archuleta, Goodrich, and Kugel 2024). Currently, two non-exclusive models have been proposed to describe the preinitiation complex assembly: the stepwise

model and the holoenzyme model. In first model, the factors are recruited to the promoted region sequentially: TFIID/TFIIA, TFIIB, RNA Pol II/TFIIF, TFIIE, and finally the TFIIH which carries a CDK7 kinase subunit. The holoenzyme model proposes instead the DNA-independent assembly of the preinitiation complex which binds later to the promoter as a single unit. In both models, the recruitment of the RNA Pol II to promoter DNA occurs in its hypo-phosphorylated form, and the phosphorylation of the Ser5 residue by the CDK7 subunit of the TFIIH facilitates the transcription (Archuleta, Goodrich, and Kugel 2024; Corbett 2018). The assembly and DNA binding of the initiation complex, and thereby the transcription rate of genes, depends on the combined actions of transcriptional activators and repressors, which respectively facilitate or inhibit the RNA Pol II binding and the subsequent initiation of transcription. More precisely, activators and repressors bind to certain DNA regions known as enhancers, which contain sequence-specific binding sites for these ligands. Activators promote the binding of the RNA Pol II whereas (Archuleta, Goodrich, and Kugel 2024).

As previously mentioned, aside from the joint action of the general TFs, transcription also depends on chromatin remodeling events which can modulate gene accessibility to the transcriptional machinery, as well as a collection of epigenetic modifications which are crucial for numerous biological processes. The most studied epigenetic mark is DNA methylation, which consists of the addition of a methyl group to the 5th carbon of the cytosine base producing a 5-methylcytosine (5-mC). The 5-mC deposition is catalyzed by a group of DNA methyltransferases (DNMTs) and its consequences for gene expression depends on the genomic context (A. V. Lee, Nestler, and Chiappinelli 2024).

DNA methylation mostly occurs in cytosines preceding a guanine nucleotide (CpG). The vast majority of gene promoters contain CpG-rich regions known as CpG islands, which are well-characterized transcriptional cis-elements. DNMTs exert their gene silencing function mostly through the hypermethylation of CpG islands subsequently repressing transcription of target genes. The repression of gene expression thus occurs mostly via the recruitment of proteins recognizing 5-mC and the interference with TFs preventing transcription (A. V. Lee, Nestler, and Chiappinelli 2024; Younesian et al. 2024; Corbett 2018). However methylation in CpG islands only constitutes a small fraction of the global 5-mC compared with the rest of the genome. Most of 5-mC are located in transposable elements which account for approximately half of the human genome (A. V. Lee, Nestler, and Chiappinelli 2024).

Key biological processes involving DNA methylation include cell differentiation, maintenance of genome stability, genome imprinting and X-chromosome inactivation (A. V. Lee, Nestler,

and Chiappinelli 2024). Alterations of the methylation mechanisms can subsequently affect multiple of these biological functions. Indeed, aberrant patterns of DNA methylation are often associated with several human diseases and some of these patterns have been recently labeled as cancer signatures (Steele et al. 2022).

In this context, histone modifications are also fundamental regulatory mechanisms that potentially modulate the chromatin structure, thus altering the accessibility of genes for the transcriptional machinery (Younesian et al. 2024). Histones can be acetylated and methylated at specific residues, triggering conformational changes that subsequently affect the degree of chromatin compaction. Furthermore, there is an intricate crosstalk between histone modifications that has been reported to be dysregulated in cancer, thus awakening the general interest for these modifications as a critical factor in cancer development (Kawaf, Ramadan, and El-Awady 2024; Velázquez-Cruz et al. 2021).

Aside from previous epigenetic marks, many non-coding RNAs (ncRNAs) also contribute to the transcriptional control of gene expression, for instance, via gene silencing. In summary, these epigenetic mechanisms act as the first filter for gene expression and are essential for the preservation of the genome integrity and stability (Younesian et al. 2024; Corbett 2018).

3. Post-transcriptional regulation of gene expression

After DNA is transcribed into RNA, post-transcriptional regulation of gene expression (PTR) encompasses the whole collection of events taking place between transcription and translation that ultimately determine the fate of transcripts, representing a second key regulatory layer for gene expression (Corbett 2018).

Over the past two decades, an increasing amount of work has emphasized post-transcriptional mechanisms as modulators of protein levels and isoform ratios, being determinants of the resulting cell phenotypes. Indeed, the PTR influences all aspects of the RNA life cycle, including RNA processing steps, control of mRNA stability and turnover, transport, localization and translation (Corbett 2018).

3.1. From precursor to mature mRNA molecules: mRNA processing steps

In mammalian cells, as soon as DNA is transcribed into precursor mRNA (pre-mRNA), a series of processing steps within the nucleus are required to produce a mature mRNA, which can subsequently be translated in the cytoplasm. The nascent RNA molecule emerging from the RNA Pol II is almost immediately “capped” through the addition of a 7-methylguanosine (m^7G) to its 5’end, which plays a major role in translation initiation and protects the RNA from being degraded by exonucleases (Cockman, Anderson, and Ivanov 2020; Corbett 2018).

The pre-mRNA consists of a coding sequence (CDS) flanked by two untranslated regions (UTR) at the 5’ and 3’ ends. After transcription, the CDS of the pre-mRNA still contains non-coding regions named introns interspersed with coding regions, or exons. A critical step in mRNA processing, known as splicing, removes these intronic regions and joins the coding exons to ultimately generate the mature mRNA molecule necessary for protein synthesis(Naro, Ruta, and Sette 2024; Corbett 2018). This essential step in gene expression is mediated by the spliceosome, a macromolecular complex that assembles on specific sequence elements within the pre-mRNA, ensuring the precise excision of introns and ligation of flanking exons. The spliceosome assembly is guided by the factors U1 small nuclear ribonucleoprotein (snRNP), which recognize the GU dinucleotide at the 5’ splice site, and the SF1, U2AF2 and U2AF1 that bind to the 3’ splice site. These factors subsequently recruit the U2 snRNP and the U4/U5/U6 tri-snRNP, activating the splicing reaction (Naro, Ruta, and Sette 2024; Wigington et al. 2015).

Splicing is commonly modulated by a wide variety of additional factors which can interfere with the spliceosome activity causing intron retention or exclusion of specific exons. In this context, many RNA-binding proteins (RBPs) and ncRNAs contribute to the alternative splicing of pre-mRNAs, generating multiple transcript variants, each containing a variable set of exons. This ultimately leads to the production of different protein isoforms, increasing the proteome diversity (Corbett 2018; Wigington et al. 2015; Naro, Ruta, and Sette 2024). Changes in the expression and activity of such splicing factors can significantly influence the alternative splicing. These splicing factors are, in turn, modulated by signaling pathways, epigenetic modifications and alterations in the RNA Pol II transcription (Naro, Ruta, and Sette 2024).

While many splicing events occur co-transcriptionally, pre-mRNAs are cleaved at the 3’-end and polyadenylated after transcription is completed, resulting in a mature mRNA. mRNAs are then packaged into ribonucleoprotein particles (RNPs) and exported to the cytoplasm. This process is mediated by export receptors present in the RNP complex which can directly interact with the nuclear pore (Cui et al. 2024; Corbett 2018; Wigington et al. 2015).

3.2. The functional diversity of non coding RNAs

RNAs perform a diverse range of biological functions and can be classified into different types according to their roles. mRNAs are the intermediary molecules responsible for transferring the genetic information encoded in the DNA to the translation machinery in the cytoplasm. However, not all of the transcribed RNA is later translated into protein. Other RNA species act as trans-factors involved in numerous regulatory processes, including interference with the mRNA processing, stability and degradation as well as the translation efficiency (Gao et al. 2020). In fact, a large portion of the transcriptome consists of non-translated RNAs which serve diverse regulatory and structural functions within the cell other than encoding proteins.

The broad set of ncRNAs functions range from regulating gene expression and chromatin structure to preserving genome stability, to also affecting key post-transcriptional processes such as mRNA splicing, editing, transport and degradation (Fu 2014; Fernandes et al. 2019). Gene expression is also influenced by the activity of miRNAs which bind target mRNA molecules promoting mRNA decay or translation inhibition (Fu 2014; Gao et al. 2020).

Among ncRNA species, ribosomal RNAs (rRNAs) and transfer RNAs (tRNAs) are vital for translation. rRNAs are an integral part of the 40S and 60S ribosomal subunits in eukaryotic cells. Human ribosomes contain four types of rRNA, the 28S, 5S and 5.8S in the 60S subunit and 18S in the 40S subunit (Kumari, Groza, and Aguilo 2021). rRNAs not only provide structural support to the ribosome but also play an important role in peptide bond formation during translation. rRNAs are essential mediators of genetic decoding and translation, playing a central role in peptide bond formation, while ribosomal proteins primarily contribute to the structural integrity of the ribosome. This is achieved through folding into specific three-dimensional structures, providing docking sites for interactions with mRNAs and tRNAs (Bastide and David 2018; Catalanotto et al. 2023). The tRNAs act as adaptors between the nucleotide sequence of mRNA and the and the amino acid sequence of proteins. Each tRNA carries a specific amino acid and possesses an anticodon loop that pairs with complementary codons on the mRNA. The pairing between anticodons and amino acids ensures the correct placement of amino acids in the growing polypeptide chain. As for rRNAs, the proper folding of tRNAs is crucial for their biological role in protein assembly (Bastide and David 2018).

3.3. RNA-binding proteins: the main players of PTR.

The post-transcriptional regulatory network is composed of a myriad of trans-factors, including RBPs, long non-coding RNAs (lncRNAs) and microRNAs (miRNAs). These factors can interact with cis-elements contained in the mRNA itself (Wigington et al. 2015). Furthermore, the multiple trans-factors participating in the PTR not only interact with their respective RNA targets but also with one another, endowing this network with a high degree of complexity and interconnectivity (Quattrone and Dassi 2019).

Among the pool of regulatory elements, RBPs have been recurrently described in the literature as the main players of PTR (Hentze et al. 2018; Quattrone and Dassi 2019; L. Street et al. 2023). RBPs represent a class of approximately 2000 proteins with the unique ability to interact directly with RNA molecules (Wassmer et al. 2024; Gerstberger, Hafner, and Tuschl 2014). These RBP-RNA interactions are mostly possible due to the presence of RNA-binding domains (RBDs) and protein regions that, even though are not structured as domains, can nonetheless bind to RNA. RBPs can recognize their target RNA molecules through different modalities, including sequence specificity, RNA secondary structures and RNA modifications (**Figure 2A**).

The most common RBDs found in RBPs are the RNA recognition motif (RRM), K homology (KH) domain, double stranded RBD (dsRBD), Zinc finger (ZnF) domain and cold-shock domain (CSD) (**Figure 2B**) (Glisovic et al. 2008; Hentze et al. 2018; Castello et al. 2016). These RBDs are highly conserved across species and play key roles in RNA metabolism. The presence of such RBDs in RBPs determines their specificity. Such specificity is accomplished by carrying multiple RBDs of the same type distributed in tandem and combined with other types, providing unique RNA recognition patterns that contribute to diversifying their RNA-protein interactions (Hentze et al. 2018; Glisovic et al. 2008).

The RRM is the most prevalent and well-studied RBD, and is estimated to occur in 1% of all human proteins. It comprises approximately 90 amino acids in size forming two α-helices against an antiparallel β-sheet (i.e. $\beta_1\alpha_1\beta_2\beta_3\alpha_2\beta_4$ topology) containing two conserved RNP1 and RNP2 motifs, creating a platform for RNA binding (Corley, Burns, and Yeo 2020). These domains are often found in tandem increasing the RNA affinity and specificity, allowing recognition of longer sequences and structural elements such as stem loops in single stranded RNAs (ssRNAs) (Corley, Burns, and Yeo 2020; Hentze et al. 2018).

The KH domain is smaller than RRM (70 amino acids) and recognizes shorter sequences of only 4 nucleotides in ssRNA. KH domains adopt a $\beta_1\alpha_1\alpha_2\beta_2\beta_3\alpha_3$ topology in eukaryotes (and reversed in prokaryotes) with a conserved GXXG RNA-binding motif between $\alpha_1\alpha_2$ helices. As with RRMs, KH domains can synergistically increase the RNA-binding specificity (Corley, Burns, and Yeo 2020).

The dsRBD is the third most common RBDs. Its specificity resides on the exclusive recognition of double stranded dsRNA which confer them critical roles in viral protection and RNA interference.The dsRBDs are composed of a β-sheet flanked by two α-helices ($\alpha_1\beta_1\beta_2\beta_3\alpha_2$). The dsRBDs recognize dsRNA helical structures up to 16 base-pairs with hydrogen bonds contacting the phosphodiester backbone and 2'-OH of the nucleotide (Corley, Burns, and Yeo 2020; Hentze et al. 2018).

ZnF domains are characterized by coordination of zinc ions, which stabilize their structure and facilitate the interaction with short RNA motifs or structured elements. The ZnF domains are subdivided in different subtypes according to the interspersed positions of cysteine (C) and histidine (H) residues. The subtypes that interact with RNA include the CCHC, CCCH, CCCC and CCHH. Among these subtypes, the CCHH ZnF is the most versatile, being able to interact with ssRNA, dsRNAs and DNA (Corley, Burns, and Yeo 2020; Glisovic et al. 2008).

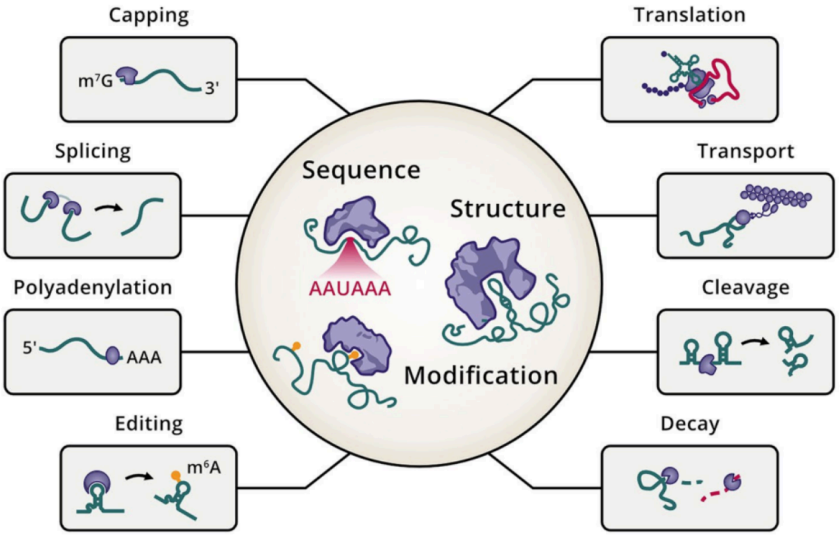
The CSD is found in a large family of proteins associated with cold adaptation. It is composed of more than 70 amino acids in eukaryotes, forming a barrel structure of five antiparallel β-strands known as the oligosaccharide/oligonucleotide-binding (OB) fold. Similarly to RRMs, the OB of CDS also contains the RNP1 and RNP2 motifs that recognize short sequences of 3-4 nucleotides in ssRNA and ssDNA. An example of RBPs carrying the CSD is the group of Y-box proteins, involved in RNA stabilization and stress response mechanisms (Corley, Burns, and Yeo 2020; Hentze et al. 2018)

While the size of the domains described above does not exceed the 100 amino acids, Helicase domains range from 350 to 400 amino acids being one of the largest RBDs known. These domains are composed of an ATP-catalytic core and a binding region coordinated by other subdomains which confer them the the capacity to unwind dsRNAs and DNA. Helicases are classified in six superfamilies (SFs), and only the members of SF1 and SF2 contain the eukaryotic RNA and DNA helicases. The RNA-binding helicases belong to the

Upf1-like, DEAD-box, DEAH, RIG-I-like Ski2-like, and NS3 families from SF1 and SF2 (Corley, Burns, and Yeo 2020; Hentze et al. 2018; Gerstberger, Hafner, and Tuschl 2014).

Aside from RBDs recognizing sequence pattern and RNA structural elements, some RBDs specifically recognize RNA modifications within the sequence. This is the case of the YTH protein family, carriers of the YTH domain, which confers the ability to recognizes N6-methyladenosine (m^6A) marks in RNA enabling selective binding to methylated over unmethylated adenosine (Corley, Burns, and Yeo 2020; Micaelli et al. 2022; Destefanis et al. 2024).

A)



B)

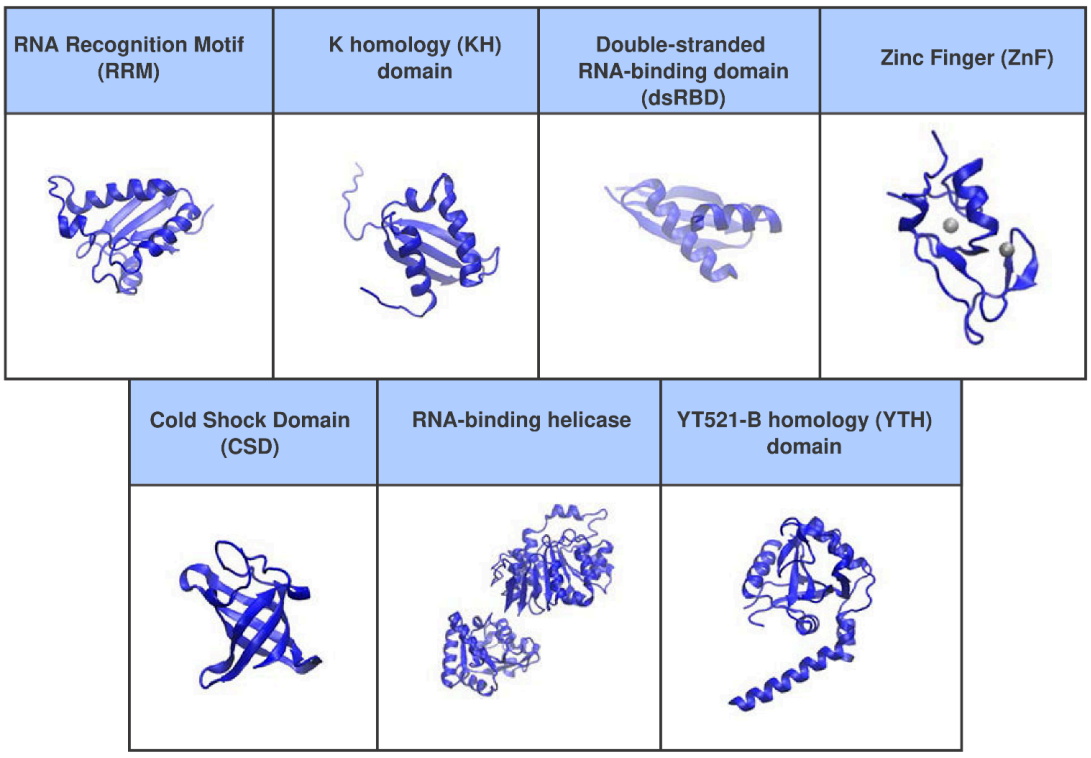


Figure 2: The different modalities of RNA-protein interactions are mediated by specific RNA-binding domains. *(**A**) RNA recognition modes of RBP and their different roles in RNA biology (image from (Gallardo-Dodd and Kutter 2024)) (**B**) three-dimensional structure of most abundant RBDs in eukaryotes (image from (Corley, Burns, and Yeo 2020)).*

Despite their high prevalence in RBPs, the RBDs described above only represent a small fraction of the total RBDs existing, and many of these RBDs remain to be

characterized(Corley, Burns, and Yeo 2020). In addition, aside from RBPs having well characterized RBDs, many non canonical RBPs lacking conventional RBDs have been described in the literature as well. Indeed, some RBPs, despite having primary roles unrelated to RNA binding, can somehow moonlight the RNA-binding function, participating in RNA processing and regulation under specific conditions. These non canonical RBP-RNA interactions are mediated by conformational changes within intrinsically disordered regions (IDRs), protein-protein interaction interfaces and enzymatic cores (Wassmer et al. 2024; Fierro-Monti 2024; Hentze et al. 2018).

In summary, RBPs are involved in all aspects of PTR, ranging from mRNA preprocessing to the control of mRNA stability, degradation and translation. The interaction between a canonical RBP and its respective pool of RNA targets is classically mediated by RBDs that can recognize specific sequence patterns of variable length. However, alternative RNA-binding mechanisms are also common on RBPs, such as the affinity for specific RNA secondary structures and the recognition of numerous chemical modifications widespread throughout the transcriptome (Zhaowei Wang et al. 2022; Destefanis et al. 2021; Navarin and Costa 2017).

3.4. Cis-element-mediated RBP-RNA interactions

RNA-binding sites of RBPs are found throughout all transcript regions. Many of these binding sites map on the 5’ and 3’ UTR flanking the CDS. RBPs binding 5’UTR and 3’UTR are often associated with different regulatory roles. On one hand, RBP binding in 5’UTR is often associated with nuclear export, cellular localization and control of translation initiation. The binding of certain RBPs at the 5’UTR can enhance translation or repress it by facilitating or obstructing the assembly of the ribosomal complex . On the other hand, RBPs targeting the 3’UTR are often involved in mRNA stability and decay (Mayr 2019; Mignone et al. 2002). For a specific group of RBPs, 3’UTR binding is mostly mediated by the recognition of cis-acting elements known as adenylate/uridylate-rich elements (AREs) (Wigington et al. 2015; C. Y. Chen and Shyu 1995). The mRNA turnover is mostly regulated by AREs, which promote mRNA decay in response to a variety of specific intra- and extra-cellular signals (Mayr 2019; Mignone et al. 2002).

The most well-studied case is the human antigen R (HuR/ELAVL1), which recognises 3’UTR AREs of their target transcripts conferring them stability. In contrast, the T-cell intracellular antigen 1 (TIA1) also recognises AREs in 3’UTR and intronic regions, but TIA1 is instead a

modulator of splicing and translation of their targets. HuR and TIA1 have been shown to be competitors upon specific mRNAs such as the PDCD4 transcript, inducing opposite outcomes over their target mRNA molecules (Wigington et al. 2015).

Cis-elements can also act as markers for a specific subset of transcript, as is the case with the 5' Terminal OligoPyrimidine (5'TOP), which consist of a repetition of 4-15 pyrimidines located immediately after the m7G cap. 5'TOPs are found in most of the transcripts coding for proteins involved in translation, such as ribosomal proteins and numerous translation factors. The translation rate of these transcripts marked with the 5'TOP is decreased under hypoxia and amino acid deficit. The inhibition of the mTOR signaling pathway also produces a decrease in the translation rate of TOP mRNAs, indicating the involvement of mTOR in the control of translation efficiency (Cockman, Anderson, and Ivanov 2020).

Another fundamental cis-element for PTR is the RNA G-quadruplex (RG4). RG4s are dynamic structures composed of four-stranded building blocks known as G-quartets, each consisting of stacks of guanine tetrads, stabilized through Hoogsteen hydrogen bonds. These cis-elements are widespread across the transcriptome and significantly accumulated in UTRs and ncRNAs, representing binding hotspots for many RBPs. The RG4 indeed exerts its role through interactions with RBPs which control the G-quartets folding. RG4s are highly involved in all post-transcriptional regulation steps and alterations of such structures are associated with cancer and neurodegenerative diseases (Bourdon et al. 2023; Dumas et al. 2021).

We have briefly described multiple modalities of RBP-RNA interactions involving different cis-elements and other molecular mechanisms, providing some examples for each of them. In this context, and because of their increasing importance, the following section is exclusively dedicated to one class of cis-elements; RNA modifications. It will thus cover the main modifications and their most well-known interactors, as well as their roles in human diseases.

3.5. RNA modifications shape gene expression: the epitranscriptome

RNA modifications play a crucial role in the regulation of most gene expression programs. While modifications of different RNA species have been reported for over five decades, their biological roles have remained mostly unknown until recently and the functionality of many of them still represents a major knowledge gap. To date, more than 170 types of RNA modifications have been identified (Kumari, Groza, and Aguilo 2021). The field of study

encompassing the landscape of dynamic biochemical marks throughout the transcriptome is known as epitranscriptomics, an emerging concept which has recently gained substantial recognition among the scientific community for its potential applications to several human diseases (Destefanis et al. 2021; Lo Giudice et al. 2020).

RNA modifications can be installed by stand-alone enzymes known as writers or by writer complexes composed of multiple subunits. Whereas some modifications are irreversible, there are many cases of reversible chemical marks whose removal is mediated by erasers (Kumari, Groza, and Aguilo 2021). Aside from the antagonic roles of writers and erasers, there is a third group of proteins with the ability to recognize specific chemical marks on their target transcripts. These proteins, called readers, are deeply involved in the RNA metabolism, and many of them have been reported as modulators of processes such as pre-mRNA preprocessing, mRNA stability and translation (H. Huang et al. 2018). In this context, the maintenance of a correct balance between writers, erasers and readers is essential for proper cellular function and overall fitness, as it allows for precise PTR and enables cells to counteract environmental changes.

The epitranscriptome is enriched with a vast array of chemical modifications that fine-tune RNA stability, translation, and function, encompassing over 170 distinct modifications which form this complex regulatory layer. These modifications impact various aspects of RNA metabolism, including stability, localization, translation, and interactions with proteins. While each modification contributes to the intricate regulation of gene expression, a comprehensive discussion of all known RNA modifications is beyond the scope of this section. Instead, we focus on a subset of well-characterized modifications that are particularly relevant to mRNA regulation and have been extensively studied in the context of cellular function and disease, namely N^6 methyladenosine (m^6A), pseudouridine (Ψ), and 5-methylcytosine (m^5C) (Nombela, Miguel-López, and Blanco 2021).

3.5.1. N^6 methyladenosine (m^6A)

N^6 methyladenosine (m^6A) is the most abundant internal (non-cap) modification in mRNA, which is estimated to be found in one-third of mammalian mRNAs, each containing around 3 to 5 m^6A sites on average. m^6A is a reversible and dynamic modification whose turnover is mediated by specific methyltransferases and demethylases (Sun et al. 2023; Gu et al. 2020) . The m^6A deposition is catalyzed by a conserved m^6A writing complex, which includes an enzymatic core, comprising a heterodimer of the methyltransferase-like 3 and 14 (METTL3

and METTL14), also known as the m^{6}A-METTL Complex (MAC). More precisely, METTL3 contains the catalytic domain that methylates the adenosine at the N6 position whereas METTL14 acts as an auxiliary co-factor which also provides structural support to METTL3 (Destefanis et al. 2024; J. Liu et al. 2014).

Moreover, the m^{6}A writing complex is also composed of additional factors that contribute to the regulation, localization and recruitment of the m^{6}A machinery to target RNA. These factors constitute the m^{6}A-METTL Associated Complex (MACOM), required for full MAC activity. The MACOM includes the mammalian pre-mRNA splicing regulator WTAP, VIRMA, RBM15/15B, and ZC3H13 (see **figure 3A**) (Bawankar et al. 2021; Destefanis et al. 2021). These factors fulfill specific functions within the writing complex, for instance, while VIRMA facilitates the recruitment of the m^{6}A complex near the stop codons and 3'UTR of mRNAs, ZC3H13 was found to stabilize the interaction between WTAP and RBM15/15B. More recently, the E3 ligase CBLL1 (also known as HAKAI) was also identified as an associated factor of MACOM, and its presence was found to contribute to maintaining m^{6}A levels in plants and human cells (Bawankar et al. 2021; Tai et al. 2022).

m^{6}A-marked transcripts can also be demethylated by erasers. These two m^{6}A erasers are the Fat mass and obesity-associated protein (FTO) and Alkb homolog 5 (ALKBH5) (see **figure 3A**) (Destefanis et al. 2021; J. Liu et al. 2014). However, the biological relevance of m^{6}A relies on their reader RBPs. Indeed, the m^{6}A-marked transcripts are recognized by a category of proteins that can determine the fate of the RNA molecule. Among these m^{6}A readers, the YT521-B homology (YTH) domain-containing proteins (YTHDFs) are noteworthy due to their determinant roles in mRNA fate (H. Huang et al. 2018). The role of YTH protein family ranges from pre-mRNA splicing and tuning of translation efficiency to mRNA degradation turnover. The most representative example is YTHDF2, which was found to selectively bind m^{6}A-marked single-stranded mRNAs and facilitate degradation, acting as a regulator of mRNA decay. The YTH protein family has the ability to recognize m^{6}A through the action of a YTH domain highly conserved in eukaryotes (X. Wang et al. 2014).

YTHDF proteins (YTHDF1, YTHDF2 and YTHDF3) are indeed m^{6}A readers that influence various aspects of RNA metabolism (**Figure 3A**). However, their functional roles remain a topic of active discussion, with ongoing debate about whether these proteins act redundantly or perform specialized functions. While recent studies suggested that the three YTHDFs redundantly promotes mRNA decay, other studies point at functional specialization (Zaccara, Ries, and Jaffrey 2019; Zou et al. 2023). In particular, YTHDF1 has been identified as an enhancer of translation and YTHDF3 in splicing and RNA nuclear export (Zou and He 2024).

While the two perspectives seem opposite, an explanation might reside in the variation in these roles depending on the cellular context and the target RNAs.

The transcript fate is often mediated by multiple m^6A readers which can have complementary or opposite roles. In this context, cooperation or competition between m^6A readers competing for the same pool of targets are not uncommon, having diverse effects over the RNA turnover (Destefanis et al. 2021). An example of these antagonist relationships is the interplay between the insulin growth factor 2 binding proteins 1 2 and 3 (IGF2BP1/2/3) and the YTHDF2. While YTHDF2 has been extensively described as a promoter of mRNA decay, IGF2BP1,2 and 3 protect m^6A-modified mRNAs from degradation (H. Huang et al. 2018).

The presence of m^6A in mammalian transcripts is associated with a highly evolutionary conserved sequence DRACH/RRACH (D corresponds to A, G or U; R corresponds to A or G; and H corresponds to A, C or U). Among its potential site locations, high-throughput sequencing has revealed that DRACH-associated m^6A is specially enriched on 3'UTR and near mRNA stop codons, but m^6A sites can be also found across CDS and within introns (Gu et al. 2020; H. Huang et al. 2018).

The m^6A modification mediates a plethora of processes essential for the RNA life cycle. For instance, recent studies have highlighted the role of m^6A writers as translation promoters of tumor suppressor genes in an oncogenic environment (Esteva-Socias and Aguilo 2024; Tai et al. 2022; Kumari, Groza, and Aguilo 2021). It is, therefore, unsurprising that imbalances on the expression of m^6A interactors lead to aberrant m^6A patterns,which could contribute to many human diseases, including several cancer types (Gu et al. 2020). For this reason, the m^6A interactome has been in the spotlight of the scientific community for many years, and a substantial number of potential therapeutic targets have already been discovered (Micaelli et al. 2022; Gu et al. 2020).

Substantial efforts have been put into the development of methods for transcriptome-wide m^6A mapping tools. In fact, recent works have accomplished the detection of m^6A at single nucleotide resolution (Cruciani et al. 2023; L. Hu et al. 2022). On a similar line, novel computational approaches have focused on the co-occurrence of m^6A with other RNA modifications, for instance m^5C (Acera Mateos et al. 2024).

3.5.2. The impact of pseudouridine in RNA biology

If we consider the whole pool of RNAs, the pseudouridine (Ψ) is the most abundant modification, being present in all types of RNA species. Different from m^6A, pseudouridylation is an irreversible modification occurring exclusively in uridine nucleosides, which consists in the isomerization of 1-ribosyluracil to 5-ribosyluracil (Ge and Yu 2013; Yu and Meier 2014; Kumari, Groza, and Aguilo 2021).

Ψ was first discovered in tRNAs, and later on rRNAs and mRNAs as well as multiple ncRNAs. In fact, Ψ is the second most abundant modification in rRNAs and is known to play important roles in protein translation. In this context, a depletion in Ψ at two specific sites in the 28S ribosomal RNA was found in patients with X-linked dyskeratosis congenita, a disease triggered by point mutations in the pseudouridine synthase gene DKC1 (Taoka et al. 2018). In contrast, to date, only some mRNA-related functions of Ψ have been reported. Some studies suggest that Ψ contribute to the mRNA stability and could also be involved in promoting translation (Destefanis et al. 2021). Ψ is predominantly facilitated by RNP complexes guided by small nucleolar RNAs (snoRNAs) containing H/ACA boxes (**Figure 3B**) (Kiss et al. 2004; Z.-H. Huang et al. 2022). In addition to DCK1, other known Ψ writers targeting mRNAs include PUS1, PUS4, PUS7 and RPUSD2 (**3B**) (Martinez et al. 2022; Levi and Arava 2021; Schwartz et al. 2014).

3.5.3. 5-Methylcytosine: Mechanisms and implications

5-methylcytosine (m^5C), similar to m^6A, is also a reversible modification with well-known writers, readers and erasers. Despite being initially discovered in rRNAs and tRNAs, m^5C has also been recently reported in mRNAs and ncRNAs, suggesting a broader distribution throughout the transcriptome than it was originally expected (Kumari, Groza, and Aguilo 2021; Bohnsack, Höbartner, and Bohnsack 2019). The m^5C deposition is catalyzed by methyltransferases containing NOL1/NOP2/SUN domains, encompassing a family of seven NSUN proteins in humans as well as the DNA methlytransferase homologue DNMT2. There are four members of the NSUN family, namely NSUN2, 5, 6, and 7, that are confirmed m^5C writers of mRNAs (see **figure 3C**) (Bohnsack, Höbartner, and Bohnsack 2019).

Whereas the role of all the m^5C interactors in cancer has not been elucidated yet, some of them are overexpressed in different cancer types. For example, there is evidence for NSUN1, NSUN2 and NSUN4 being upregulated in breast cancer (BC), which might indicate

an hypermethylation of cytosine as a potential cancer hallmark. In addition, the genomic region where the NSUN2 gene is located is associated with a higher predisposition of BC development suggesting a dysregulation of this gene as a potential trigger (Frye et al. 2010; Kumari, Groza, and Aguilo 2021). Similarly to m^6A, m^5C modifications can be dynamically reversed by specific demethylases. Among these m^5C erasers, TET1, TET2 and ALKBH1 have been confirmed as key regulators of mRNAs.

m^5C also has a diverse set of readers including Y-box binding proteins (YBXs) such as, YBX1, YBX2 and YBX3, as well as other factors like ALYREF and SRSF2 (**Figure 3C**). Notably, YBX1 has been shown to stabilize the oncogene HDGF in urothelial carcinoma and is overexpressed in BC suggesting a pivotal role in cancer progression (Kumari, Groza, and Aguilo 2021).

3.5.4. RNA editing

Epitranscriptomics also encompasses post-transcriptional mechanisms of base substitution and insertion/deletions (indels) within the RNA sequence. RNA editing is defined as site-specific alterations of the RNA sequence excluding changes due to processes such as splicing and polyadenylation (Van Norden et al. 2024; Lo Giudice et al. 2020; Gott and Emeson 2000). The first RNA editing event in mammals was discovered in 1989 on transcripts of ApoB with reported cytosine deamination at the 6666 position, resulting in an early stop codon and the subsequent truncated protein with its function altered. (Lo Giudice et al. 2020; Wolfe, Arnold, and Chen 2019; Driscoll et al. 1989).

There are two families of deaminases responsible for single nucleotide substitutions; The ADAR (adenosine deaminase acting on RNA) and the AID/APOBEC (Activation Induced Deaminase/"Apolipoprotein B mRNA editing enzyme, catalytic polypeptide") family. ADARs mediate the substitution of adenosine to the guanosine-analog inosine (A-to-I) whereas APOBECs catalyze the cytosine to uracil substitution (C-to-U) (Van Norden et al. 2024; Destefanis et al. 2021; Lo Giudice et al. 2020).

A-to-I editing in dsRNA represents the most prevalent type of editing in mammalian cells and is catalyzed by ADAR1 (also known as ADAR) and ADAR2 (ADARB1) (**Figure 3D**) (Tang et al. 2020). Dysregulation of these enzymes has been implicated in cancer and neurological disorders (Tang et al. 2020; Gassner et al. 2021). A-to-I editing can alter splicing regulatory

elements, leading to aberrant splicing in various cancers, while also affecting mRNA stability, localization, translation and interactions with post-transcriptional regulatory elements (Fumagalli et al. 2015; Gannon et al. 2018; Gassner et al. 2021). Notably, specific RNA-editing loci have been identified as differentially expressed in genes associated with cancer-related pathways and linked to poorer survival outcomes (Hwang et al. 2019).

While APOBEC enzymes share a basic structural homology, they possess a diverse range of functions, differing on their impact on human health and diseases (Salter, Bennett, and Smith 2016). APOBEC1 (A1) was the first APOBEC enzyme to be characterized, and is responsible for the C-to-U editing in apoB mRNA. In humans, A1 expression is confined to the small intestine whereas in mice it is broadly expressed. (Pecori et al. 2022; Salter, Bennett, and Smith 2016). A1 is inactive in the cytoplasm and performs its RNA editing activity once it is translocated to the nucleus along with the A1 cofactor (A1CF), constituting the editosome. However, it has been demonstrated that A1 activity is actually conferred by the cofactor RNA binding motif protein 47 (RBM47), being the minimal unit for RNA editing (Wolfe, Arnold, and Chen 2019; Salter, Bennett, and Smith 2016).

APOBEC3 (A3) is the most diversified C-to-U enzyme family in humans, composed of seven different proteins (Pecori et al. 2022). While one of their main roles is the restriction of viruses and transposon elements, certain A3 members also target the host genome, causing mutations that represent a potential cause of DNA breaks and genome instability. These genotoxic effects, in turn, contribute to heterogeneity in a tumorigenic context and may drive cancer evolution. In fact, it has been shown that the overexpression of A3 genes significantly triggers carcinogenesis. Among all A3 proteins, the APOBEC3A (A3A) is the major driver of endogenous sources of mutation in cancer (Van Norden et al. 2024; Petljak et al. 2022; Pecori et al. 2022). In contrast to A3 subfamily, APOBECs 2 and 4 are the less studied due to the lack of phenotype associated with their expression, and their possible biological roles remain thus undiscovered (Pecori et al. 2022). So far, C-to-U activity on mRNA targets in mammalian cells has been confirmed for A1, A3A and A3G, being the last one context-dependent, for example under certain stress conditions such as hypoxia (**Figure 3D**)
.

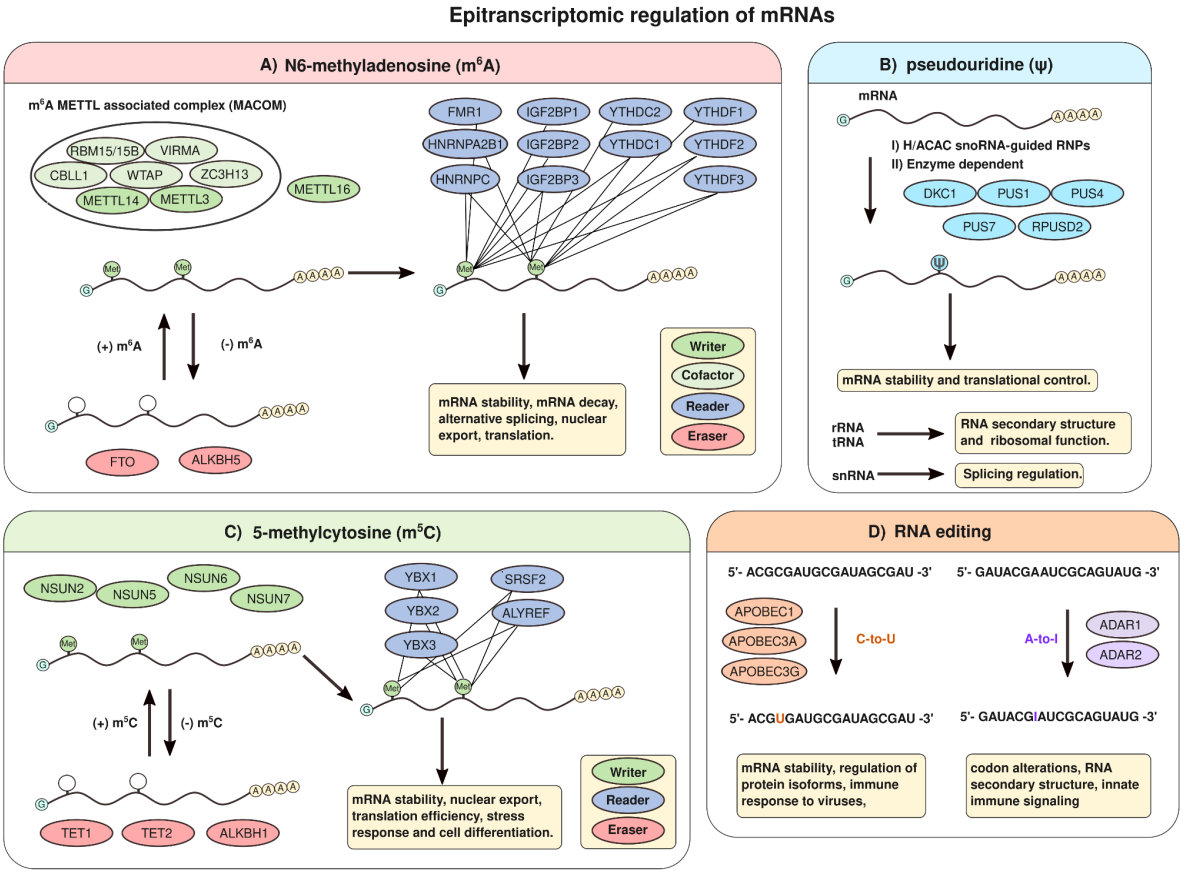


Figure 3: Epitranscriptomic regulation of mRNA: Main modifications and editing mechanisms. *Illustration of the key epitranscriptomic marks and RNA editing events that regulate mRNA function. Each panel highlights the following modifications: (**A**) N6-methyladenosine (m^6A) is installed by the methyltransferases METTL3 and METTL4 of the MACOM, removed by the demethylases FTO and ALKBH5, and recognized by reader proteins such as the members of the YTH family. m^6A is involved in mRNA stability, degradation, translation and splicing. (**B**) Pseudouridine (Ψ) is an irreversible modification formed through the pseudouridine synthases (PUS1, PUS4, PUS7 and RPUSD2) or by the H/ACA snoRNA-guided RNPs (e.g. DKC1). Ψ enhances RNA stability, translation and splicing. (**C**) 5-methylcytosine (m^5C) in mRNAs is catalyzed by members of the NSUN family (NSUN2/5/6/7) and removed by TET1/2 and ALKBH1 enzymes. m^5C affects mRNA stability, nuclear export and translation efficiency. (**D**) RNA editing: APOBEC-mediated C-to-U editing consists in the conversion of cytidine (C) to uridine (U), which can produce protein isoforms and affect mRNA stability. ADAR-mediated A-to-I editing modifies adenosine (A) to inosine (I) in dsRNA regions impacting codon recoding and splicing.*

3.6. Post-translational modifications (PTMs) modulate RBP activity

RBP activity is influenced by many different factors ranging from oscillations of pH and temperature, to biochemical reactions, such as the addition/removal of functional groups. Post-translational modifications (PTMs) are a set of covalent processing events occurring after protein biosynthesis which affect protein properties and diversify the range of protein functions. PTM deposition can lead to conformational changes of the protein which, in the case of RBPs, can modulate their RNA-binding affinity (Velázquez-Cruz et al. 2021).

Even more interesting is the existence of PTM crosstalks. An increasing amount of work evidence the fact that PTMs can potentially interfere or facilitate the deposition of other PTMs, adding an extra layer of complexity to the regulation gene expression. Indeed, a recent study revealed positive and negative correlation between acetylation and phosphorylation deposition in specific proteins; its ratio was altered across many different cancer types (Geffen et al. 2023). This fact suggests a crosstalk between kinases and acetyltransferases which is somehow disrupted under certain conditions.

PTM dysregulation has been reported as a potential cause of main human diseases, often acting as triggers of different types of cancer and neurodegenerative disorders such as amyotrophic lateral sclerosis (ALS) and frontotemporal dementia (FTD) (Sternburg, Gruijs da Silva, and Dormann 2022). However, whether aberrant PTM patterns should be considered cause, consequence or passengers, is still uncertain for many diseases. Recent studies highlight the importance of PTMs in cancer development and progression. An example of this is the dysregulation of SUMOylation and phosphorylation deposition on the RBP HuR has been shown to trigger liver and colon cancer respectively (Lachiondo-Ortega et al. 2024; Velázquez-Cruz et al. 2021).

4. Overview of cross-linking and immunoprecipitation (CLIP) methods

Exploring the different roles of RBPs in PTR requires experimental techniques to identify the set of RNAs that they bind. In this regard, two approaches can be distinguished according to the purpose of the assay. While the protein-centric approaches aim to identify the set of RNA sites for a specific RBP, the RNA-centric approaches instead consist of detecting the subset of proteins binding a specific transcript, or set of transcripts of interest (Hafner et al. 2021).

The most widely used protein-centric methods up to date are based on RNA immunoprecipitation (RIP) and crosslinking and immunoprecipitation (CLIP). CLIP techniques rely on the irradiation of cells with UV light, causing covalent bonds between proteins and their pool of RNA substrates, i.e. those RNA molecules in immediate vicinity at the moment of the UV irradiation (Hafner et al. 2021). Despite the many variations of CLIP, its core principles mostly remain the same. The UV irradiation generates ribonucleoprotein (RNP) complexes where RNA stretches covered by an RBP remain protected from degradation by the subsequent RNase treatment, allowing the isolation of those RNA fragments. Then, the RNPs of interest are selected using immunoprecipitation and the protein-bound RNA is further purified by SDS-PAGE. Finally, the sequencing of RNA fragments and the computational preprocessing steps allow the detection of the pool of RNA-binding sites for a specific RBP (**Figure 4A**) (Hafner et al. 2021; Van Nostrand et al. 2016a).

The original CLIP technique has several potential sources of bias, including issues with RNA-protein crosslinking efficiency, library preparation, and background noise. Over time, these limitations have been addressed with different approaches. In this context, different CLIP variations have been developed, namely individual nucleotide resolution CLIP (iCLIP), infrared (irCLIP) and enhanced CLIP (eCLIP), each offering improvements in different aspects of the procedure (Hafner et al. 2021; Van Nostrand et al. 2016a). For example, iCLIP enables single-nucleotide resolution by improving cDNA library preparation and incorporating ligation-based methods for capturing cross-linked RNA-protein complexes. irCLIP, instead leverages infrared light to enhance the crosslinking efficiency and minimize background signal, providing clearer RNA-protein interactions (Hafner et al. 2021; F. C. Y. Lee and Ule 2018; Zarnegar et al. 2016).

The most recent variant, eCLIP, significantly refines the technique by improving the capture and identification of RBP binding sites. eCLIP exhibits significant advantages over previous methods, including higher sensitivity, reduced background noise, and improved resolution, making it particularly suitable for identifying RNA-binding sites with greater accuracy (Van Nostrand et al. 2016b).

In summary, eCLIP provides more precise and comprehensive data for the analysis of RBP binding sites, representing a powerful tool for investigating the regulatory roles of RBPs in multiple biological processes. Therefore, in the present work we focused on the eCLIP method, providing an overview in the following section.

4.1. Leveraging eCLIP to address the transcriptome-wide binding profile of RBP

Among CLIP derived techniques, eCLIP presents suitable features for the computational analysis of RBP RNA-binding sites. eCLIP can be considered an evolution of the iCLIP technique since both share a common backbone of main steps with slight differences. However, eCLIP incorporates a number of modifications at different steps which constitute significant improvements in accuracy and efficiency compared to its predecessor.

Briefly, eCLIP modifications significantly reduce previous sources of technical bias on the library preparation step and the reduction of background noise. While the circular ligation step of iCLIP is quite inefficient, this problem is solved in eCLIP by adding adapters at two different steps: the 3′ RNA adapter is ligated during the immunoprecipitation, whereas the 3′ ssDNA adapter is ligated after the reverse transcription (Hafner et al. 2021; Van Nostrand, Pratt, et al. 2020a; Van Nostrand et al. 2016a).

eCLIP approach to filter background noise consists in using a size-matched input (SMI) which is generated performing the same steps as the samples except for the immunoprecipitation. The samples are normalized against the SMI in downstream steps of the computational pipeline to remove peaks derived from the background noise. This method allows to drop non-specific interactions retaining only the RNA-binding sites of the RBP of interest (Hafner et al. 2021; Van Nostrand et al. 2016a).

The eCLIP preprocessing steps are designed to ensure the accurate identification of the RBP binding sites with high quality and resolution. The pipeline starts with a quality control over the raw sequencing reads to check for overall sequencing quality and data integrity. This step includes the trimming of adapter sequences and the filtering of low quality reads (**Figure 4B**).

The next step is the removal of repeated elements such as transposon or satellite DNA, to reduce the noise and avoid artifacts produced by non-specific binding or crosslinking. The importance of this step resides in the high accumulation of reads aligning to repetitive sequences that potentially skew the interpretation of binding sites (Van Nostrand et al.

2016a). The subset of high-quality reads are then aligned to a reference genome. Accurate mapping is essential for identifying the location of RNA-binding sites. The aligned reads accumulate at specific locations constituting peaks (**Figure 4B**) (Van Nostrand et al. 2016a).

The peak calling step consists in identifying significant accumulation of peaks to discriminate between reliable RNA-binding sites and background noise. In this step, the eCLIP samples are compared against the SMI to discard false peaks due to high abundance of non-specific interactions (**Figure 4B**). The SMI normalization consists in the calculation of a fold enrichment and the corresponding statistical test to account for biases in RNA abundance. Real peaks are considered statistically significant using the false discovery rate (FDR) threshold (Van Nostrand et al. 2016a; Van Nostrand, Freese, et al. 2020). The final output is the set of peaks which are significant accumulation of reads indicative of a high RBP occupancy (Van Nostrand et al. 2016a).

Due to its reasonable peak reliability and data availability, in the present work we leveraged eCLIP assays from ENCODE database (Luo et al. 2020). Multiple downstream analyses derived from the RNA-binding information can be conducted after obtaining such peaks of RBP occupancy, for example, the discovery of binding motifs in the RNA sequence or the assessment of potential preferences for specific RNA biotypes and transcript regions (**Figure 4C**). Moreover, the availability of binding site information has also facilitated the the study of RBP-RBP interplay raising novel approaches for the prediction of such molecular interactions (Khoroshkin et al. 2024; L. A. Street et al. 2024)

Figure 4: eCLIP assay overview. *(**A**) (image from (Van Nostrand et al. 2016b)) Graphical abstract of main experimental steps described on the eCLIP protocol. (**B**) The step-by-step computational workflow for processing of eCLIP sequencing data prior to the analysis. The output is the tabular data (usually formatted as BED6) containing the set of significant peaks with high RBP occupancy which can be considered high-confidence RBP-binding sites (**C**) Downstream analysis of eCLIP output peaks. (image modified from (Hafner et al. 2021)).*

Aim of the thesis

This project first aims at the identification and evaluation of relevant RBP-RBP interaction features that can allow discriminating between competitive and cooperative behaviors. The second objective is to evaluate the impact of post-translational modifications on RBPs activity, and more precisely their effect on RBP-RBP interactions. Finally, the third objective consists in evaluating the differential expression of RNA modifying proteins using RNA-Seq data from breast cancer samples to identify gene signatures associated with the epitranscriptome.

Chapter 1: Identifying the determinants of cooperation and competition between RBPs.

Background

Transcripts are most often targeted by more than one RBP, which consequently may lead to RBP-RBP interactions (RRIs), which can be determinant for gene expression. In fact, many of these interactions have been reported in the literature, suggesting the high complexity of the post-transcriptional network. In this context, RBPs can interact with one another following different modalities (Dassi 2017; L. Street et al. 2023; Khoroshkin et al. 2024).

Two RBPs interact in a cooperative manner when their function is mediated by the simultaneous presence of both proteins on a given transcript. While cooperation can take place through physical protein-protein interactions (PPIs), it does not always imply direct contact between the two proteins. This is the case for RBPs sharing a common pool of RNAs which undergo a similar fate after being targeted by any of the RBPs (Dassi 2017).

Analogously to TFs, after binding their target RNAs, RBPs can facilitate the binding of a second RBP (Berenson et al. 2023; Dassi 2017). This recruitment is often mediated by specific PPI domains and regions such as the Src homology 3 (SH3) and the bromo domain (Berenson et al. 2023).

For instance, sequential binding among RBPs is a crucial mechanism of regulation of RNA splicing. More precisely, the recruitment of splicing factors is a necessary step for the proper spliceosome assembly, which determines the splice site recognition, influencing the alternative splicing outcome (Fredericks et al. 2015). The recruitment of RBPs can also be mediated by changes of the RNA secondary structures promoted by the recruiting RBP (Y. Fang et al. 2024).

Aside from events of sequential binding, some RBP pairs need both proteins to simultaneously bind the RNA. The cooperation is thus said to be mutual when both members are required to be present for the co-binding with the RNA molecule (Berenson et al. 2023).

In contrast to cooperation, RBPs can also compete for a common set of targets, potentially having antagonistic roles on the regulation of specific transcripts. RBP-RBP competition refers to the “contest” for binding to one or more overlapping RNA-binding sites, displacing or preventing the binding of the “rival” protein (Berenson et al. 2023; Quattrone and Dassi 2019; Dassi 2017).

For instance, RBP competition has been reported for several heteronuclear RNPs (hnRNPs) involved in mRNA processing. Two groups of RBPs, the group of heterogeneous nuclear ribonucleoproteins (hnRNPs) and the Serine and Arginine-rich (SR) family, perform pivotal roles on splicing. Competition between hnRNP and SR proteins for shared binding targets have been extensively documented in the literature. This competition often disrupts the splicing of the targeted proteins. More precisely, while SR proteins promote splicing through binding to splicing enhancer elements, some hnRNPs recognize splicing silencers thus favoring exon inclusion (Dassi 2017; Zhu, Mayeda, and Krainer 2001). For instance, hnRNPA1 counteracts the splicing activity of SF2/ASF and SC35, interfering with their ability to regulate splicing and leading to the generation of alternative transcript variants (Y. Fang et al. 2024). Also, SRSF1 prevents HNRNPA1 binding to splicing silencers, allowing the exon removal. The balance between competitive RBPs can indeed determine the excision and retention of specific exons, contributing to the generation of multiple transcripts of a single gene via alternative splicing (Y. Fang et al. 2024; Dassi 2017).

RBP-RBP competition can also take place over mature mRNA molecules. Some of these interactions are fundamental for the maintenance of the mRNA turnover rate. In this context, the mRNA stability and decay of specific transcripts can depend on which is the predominant RBP. For example, the YTHDF proteins, and in particular YTHDF2, are promoters of RNA

decay on m^{6}A-marked mRNAs. In contrast, IGF2BPs can bind to the same mRNAs but are instead enhancers of mRNA stability (Destefanis et al. 2021; H. Huang et al. 2018; X. Wang et al. 2014).

RBP competition can also affect translation regulation by promoting or delaying the assembly of ribosomal complexes. For example, interactions between G3BP1, PUM1, DDX3X and UCHL5 and other multifunctional RBPs are highly associated with cardiac translational regulation, shaping the target mRNAs translation efficiency (Schneider-Lunitz et al. 2021; Nag et al. 2022).

Antagonism between RBPs can also be decided by sequestration, preventing one RBP to bind its respective set of targets (Berenson et al. 2023). For instance, the RBP ELAVL1 (HuR) recognizes AU-rich elements in the UTRs of its target mRNAs. However a second RBP, TTP, can bind to the same pool of targets and promote RNA degradation preventing the binding of HuR (Zhonghua Wang, Bhattacharya, and Ivanov 2015). Similarly, a competitive relation involving target sequestration has been also reported for FUS and TDP in neurodegenerative diseases (Y. Liu et al. 2022; Zhonghua Wang, Bhattacharya, and Ivanov 2015).

Recent studies have demonstrated the potential of machine learning approaches to accurately predict RBP-RBP interactions (Khoroshkin et al. 2023; L. Street et al. 2023). Despite predicting potential interactors, these models do not make distinction between cooperative and competitive interactions but rather identify the potential interactors. Therefore, our work attempts at addressing this issue. While RBPs are key players in post-transcriptional regulation of gene expression there are examples of interplay involving other post-transcriptional factors (Lachiondo-Ortega et al. 2024; Srikantan, Tominaga, and Gorospe 2012). Here, our work will cover the RBP-RBP interactions leaving aside other types of interactions.

Results

Addressing cooperation and competition among RBPs.

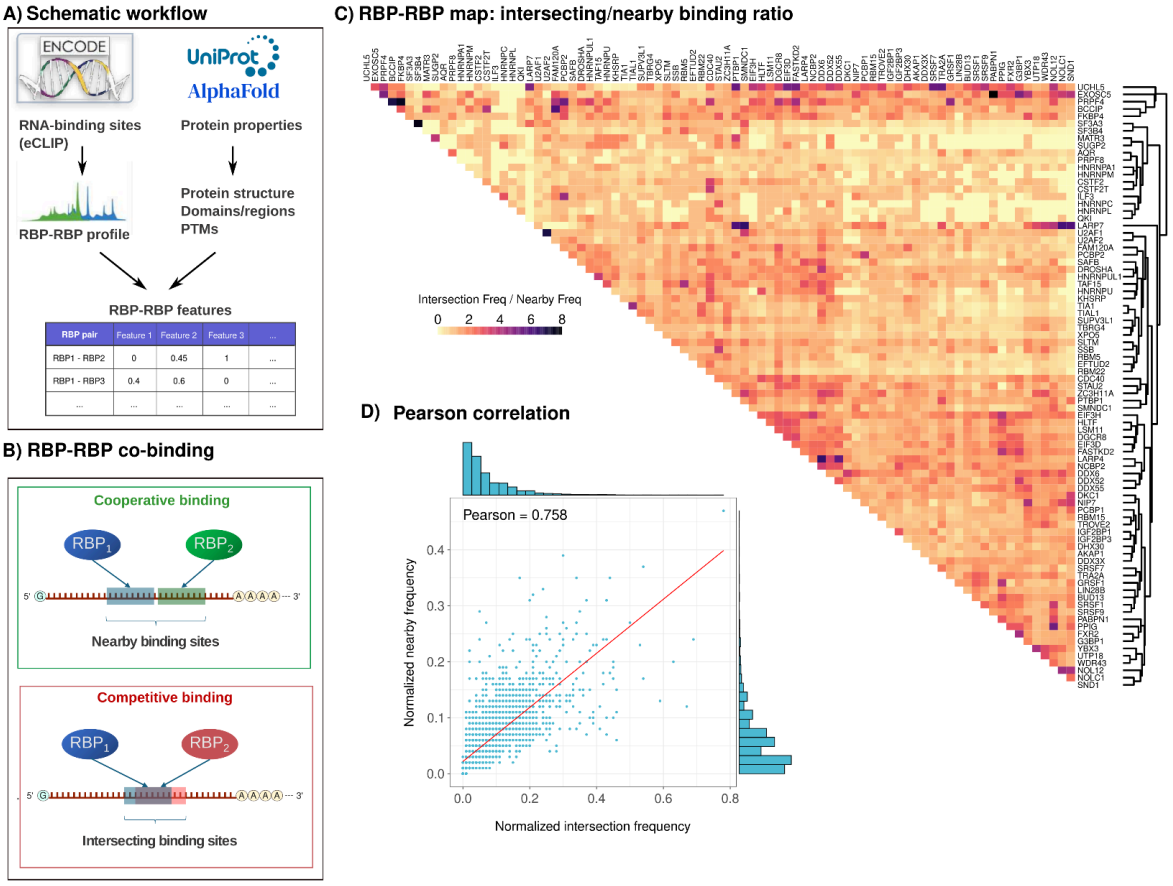
The first step towards evaluating competition and cooperation among RBPs was to gather the largest possible collection of RBP targeting data, combining information from different online sources. The main selection criteria for RBPs included in this analysis was the

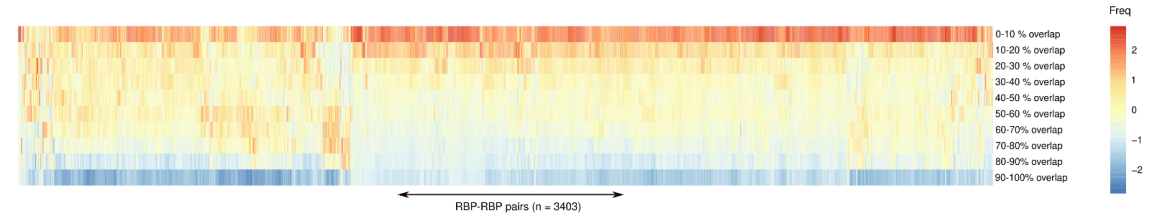
availability of eCLIP assays on ENCODE as source of RNA binding site information, in particular eCLIP experiments derived from the HepG2 cell line (Luo et al. 2020). By extracting these datasets, we obtained 83 RBPs, which were used to compute all possible RBP pairwise combinations obtaining a total of 3403. Once the data was collected, a number of preprocessing steps were performed to compute and filter a set of features used during the analysis, eventually building a matrix of features for each RBP-RBP pair (**Figure 1.1A**). This feature matrix was in turn used as input to evaluate cooperation and competition between RBP-RBP pairs. A detailed description of the databases and the computation of features is provided in the **Material & methods** section of this chapter.

The most important features for this analysis were the frequencies of nearby and overlapping binding sites obtained from overlapping profiles of RNA-binding sites of individual RBPs (**Figure 1.1B**). At the beginning, a negative correlation between nearby and overlapping binding sites was expected, which would be a sign of polarization of RBP pairs towards cooperative and competitive profiles. However, after computing Pearson correlation, we observed that frequencies of intersections and nearby binding sites were actually positively correlated (R = 0.758) (**Figure 1.1D**).

We thus opted for alternative approaches to assess differences between RBP-RBP profiles. The initial method consisted in integrating both features into a single metric by calculating the ratio of intersecting to nearby binding sites. This ratio serves as an indicator of whether an RBP pair tends to have overlapping binding sites more frequently than adjacent, non-overlapping ones (high score values) and vice versa (low score values) (**Figure 1.1C**). While this ratio could be informative, the whole set of intersections was used for the calculation regardless of the degree of overlap between binding sites of both RBPs, which could decrease the sensitivity.Similarly, the ratio neither makes distinction between long and short distances for adjacent sites.

Taking this into account, we further splitted the intersections and nearby into subsets according to the degree of overlap (from 0 to 100 percent) and the gap length between adjacent binding sites. On one hand, for the intersection frequency split, the segregation and subsequent hierarchical clustering showed a small group of RBP pairs with high-overlap intersections. These results suggest the existence of RBP pairs with a tendency to compete for the same RNA-binding sites (**Figure 1.1E**). On the other hand, the segregation of nearby RNA-binding sites by distance revealed two main groups of RBP pairs with similar size; those with predominance of short gaps and those at relatively long distance in terms of nucleotides (**Figure 1.1F**).


A) Schematic workflow
ENCODE
UniProt
AlphaFold
RNA-binding sites (eCLIP)
Protein properties
RBP-RBP profile
Protein structure
Domains/regions
PTMs
RBP-RBP features
RBP pair
Feature 1
Feature 2
Feature 3
RBP1 - RBP2
RBP1 - RBP3
B) RBP-RBP co-binding
Cooperative binding
Nearby binding sites
Competitive binding
Intersecting binding sites
C) RBP-RBP map: intersecting/nearby binding ratio
Intersection Freq / Nearby Freq
D) Pearson correlation
Pearson = 0.758
Normalized nearby frequency
Normalized intersection frequency


E) Intersecting-binding frequency splitted by overlap%
0-10 % overlap
10-20 % overlap
20-30 % overlap
30-40 % overlap
40-50 % overlap
50-60 % overlap
60-70% overlap
70-80% overlap
80-90% overlap
90-100% overlap
Freq
RBP-RBP pairs (n = 3403)

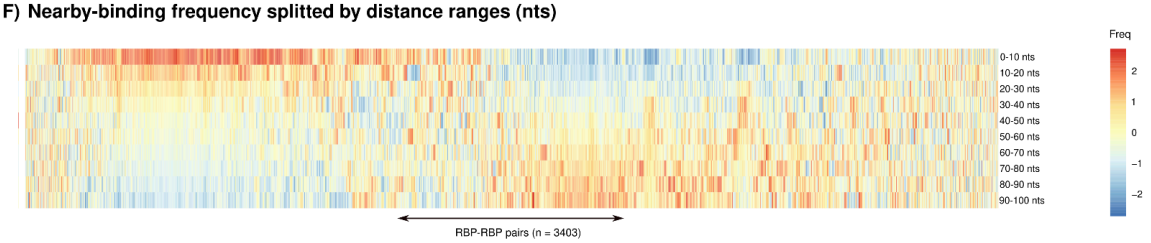

F) Nearby-binding frequency splitted by distance ranges (nts)
0-10 nts
10-20 nts
20-30 nts
30-40 nts
40-50 nts
50-60 nts
60-70 nts
70-80 nts
80-90 nts
90-100 nts
Freq
RBP-RBP pairs (n = 3403)

***Figure 1.1**: **Evaluation of intersecting and proximal RBP binding sites as features of RBP-RBP interactions** (**A**) Schematic representation of data collection and computation of RBP-RBP features. (**B**) Representation of potential cooperative and competitive events (upper and lower panels respectively) co-occurring between two hypothetical RBP pairs. Cooperative RBP pairs are expected to co-bind at nearby sites whereas competitive RBPs have a predominance of common or overlapping sites over their shared targets. (**C**) RBP-RBP heatmap. Columns and rows represent the same number of RBPs following hierarchical order. Each cell represents the ratio of intersecting/nearby binding sites between two RBPs (match column-row) (**D**) Linear correlation between frequencies of intersecting and nearby RNA-binding sites of RBP couples. Each point represents an RBP pair from the 3403 pairwise combinations from HepG2 eCLIPs. (**E,F**) Heatmaps of RBP pairs by segregation of features. Each column represents an RBP-RBP pair and the rows are the different features. The upper heatmap (E) displays the frequencies obtained for the intersections segregated by range of overlap % from 0 to 100% and the lower heatmap (F) displays the frequencies of nearby sites segregated by ranges of distances between RNA-binding sites.*

The degree of overlap and distance between RBP binding sites revealed potential RBP-RBP clusters.

The hierarchical clustering suggested the existence of subgroups of RBP pairs differing in their frequencies of intersections and proximal binding sites (**Figure 1.1E,F**). To address the existence of potential groups of interactions we opted for an unsupervised learning approach. Specifically, we applied Non-negative Matrix Factorization (NMF), a dimensionality reduction technique suitable for discovering intrinsic patterns in high-dimensional datasets. NMF consists in factorizing a non-negative matrix into two lower-rank matrices of K x W and H x K dimensions, which helps to reveal latent components and the association of each observation to these components (Rauluseviciute et al. 2023; Gaujoux and Seoighe 2010). In fact, the most important parameter to take into account in NMF is precisely the number of clusters or factorization rank, which is required to be set by the user.

Therefore, to determine the optimal number of clusters, a preliminary NMF rank survey was conducted using multiple rank estimation methods. The analysis revealed some discrepancies among the different rank estimators and consensus matrices, highlighting the complexity of cluster identification in this dataset. While the cophenetic correlation suggests

stabilization around the rank 5 (**Figure 1.2A**), further inspection of the consensus matrices revealed greater stability at rank 4 compared to rank 5 (**Figure 1.2B**). These discrepancies emphasize the importance of assessing multiple evaluation metrics in rank selection. After careful consideration of these parameters, we selected rank 4 as the optimal value for NMF clustering, ensuring an optimal cluster resolution and interpretability.

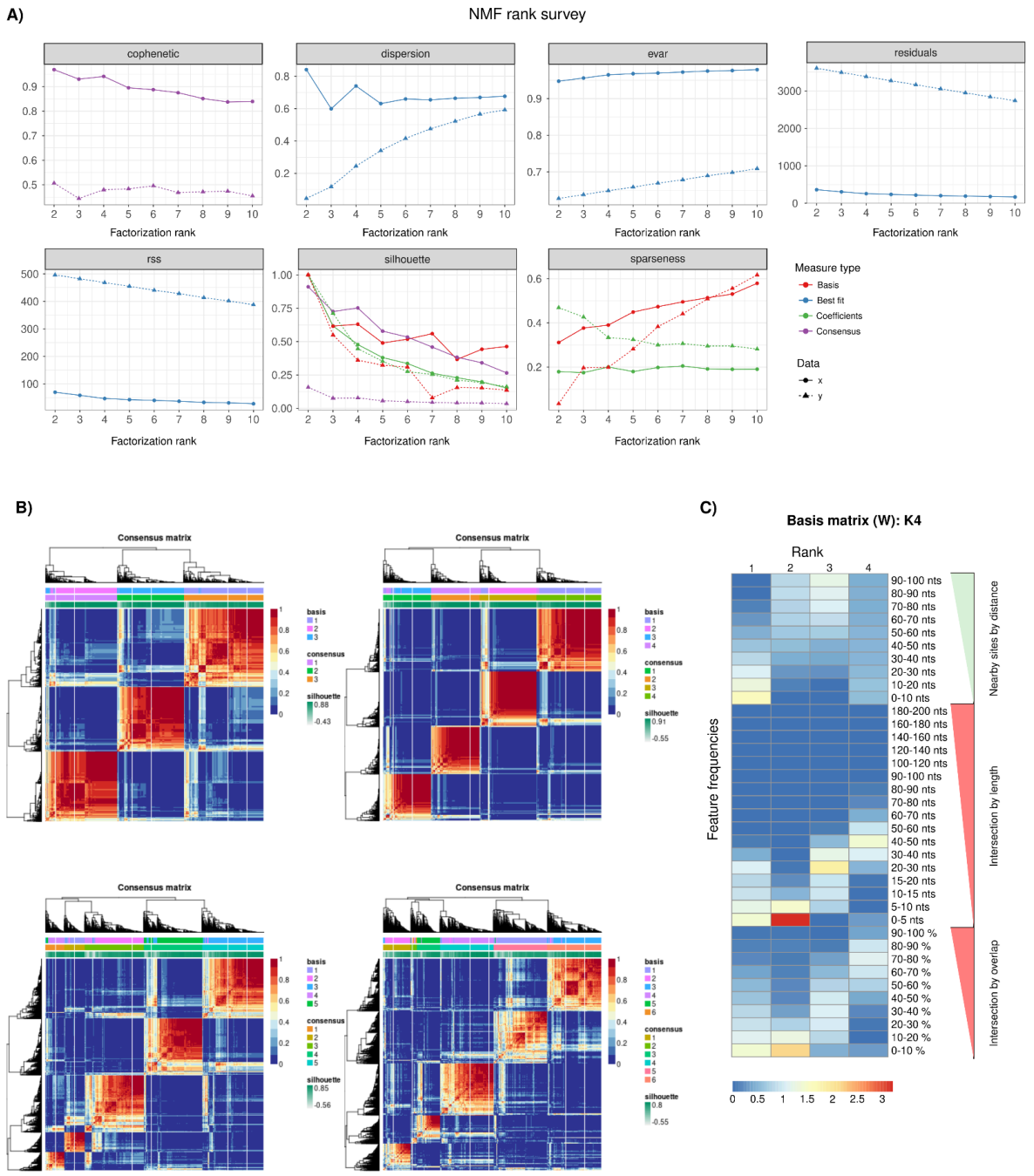


Figure 1.2: NMF rank survey and consensus matrices for segregated intersection and nearby RBP binding site frequencies. *(**A**) NMF Rank survey for the range 2-10 ranks.*

*Each line plot represents a different parameter for evaluation of optimal nº of ranks. The line of circular points display the results for the original non-negative matrix whereas the triangular points were obtained from the randomized matrix (**B**) Consensus matrices for ranks 3 to 6 . (**C**) NMF Basis matrix (W matrix) obtained for nº of factorization ranks = 4*. The columns are the four different cluster obtained for and the rows represent the segregated features indicated as three blocks: intersection frequency segregated by overlap %, intersection frequency segregated by length (nts) and nearby binding sites segregated by distance (nts).

After determining the optimal number of clusters (K = 4), NMF of rank 4 was performed over the matrix of segregated features. Each RBP pair was then assigned to the cluster with the maximum value on the coefficient matrix (H). The same procedure was followed for the features displayed on the basis matrix (**Figure 1.2C**). We thus consider the cluster assignment to split the data into four blocks corresponding to the different clusters (**Figure 1.3**).

Cluster 1 is characterized by relatively high frequencies of short nearby distances, mostly between 0 and 10 nucleotides of gap and abundant short intersections (0-10% overlap). The results suggest a tendency of RBP pairs in cluster 1 toward nearby binding sites. In contrast, RBP pairs in cluster 2 exhibit longer intersecting sections, primarily in the range of 40-50 nts, with fewer cases extending to 50-60 nts. Most intersections fall within the 40-50 % interval, with smaller proportions ranging from from 50 to 90%. Cluster 3 exhibits an accumulation of intersections ranging from 0 to 10 % overlap and short lengths spanning from 0 to 5 nts which, similarly to cluster 1, suggest close non-interfering sites. Cluster 4 shows a high frequency of long intersections (20-30 and 30-40 nts) similar to cluster 2 and a slight tendency to overlaps ranging from 40 to 50% (**Figure 1.3**).

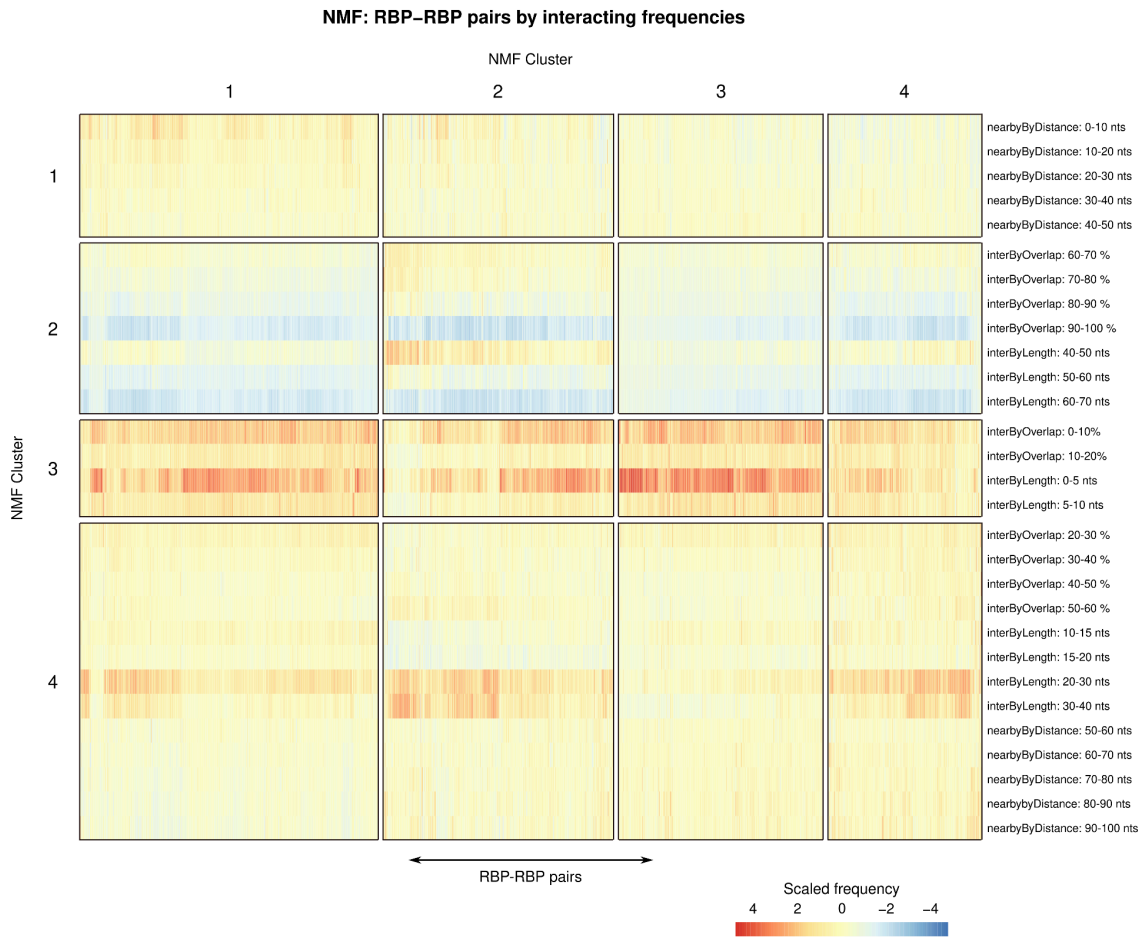


Figure 1.3: RBP pairs clustering based on the frequency of segregated interaction features. *The columns represent all 3403 RBP pairwise combinations from 83 ENCODE RBPs with eCLIP data (HepG2 cell line). Each row represents a feature obtained from the segregation of frequencies into multiple ranges. Nearby binding sites are stratified by distance (nearbyByDistance) and intersections by intersection length (interByLength) and overlap % (interByOverlap). The matrix is splitted according to the NMF clusters identified in previous steps.*

Discussion

In an initial attempt to classify RBP couples, we leveraged eCLIP assays to identify RNA-binding sites of both RBPs and evaluate their behaviour based on the positional information. We investigated the extent of overlap between binding sites of two RBPs and examined how main sites were located within close proximity, defining a maximum gap of

100 nucleotides between sites. This threshold for proximity was arbitrarily set at 100 nts to provide a delimit working definition; however, the actual spatial distance between binding sites depends significantly on the folding architecture of the RNA molecule. For instance, two pairs of RNA-binding sites in separate transcript with identical nucleotide distance may exhibit different spatial proximities due to differences in RNA secondary structures. Transcripts with a higher degree of folding will present shorter effective distances between binding sites. Although the current approach does not incorporate RNA secondary structure into the analysis, we expect these new features to provide more accurate insights into binding site real proximities in the future.

Building on these observations, we next focused on quantifying the frequencies of intersecting and nearby binding sites for each RBP pair to further characterize their interaction patterns. In retrospective, higher frequencies of intersecting binding sites were expected to be indicative of competitive RBP pairs, whereas nearby sites should be the predominant event in cooperative pairs. However, a strong positive correlation between frequencies of intersection and nearby sites was observed (**Figure 1.1D**). This result could suggest that both features reflect the frequency of RBP-RNA interactions rather than providing insight into the nature or mechanisms of RBP-RBP interactions. However, the computation of the intersection/nearby binding sites ratio as a score for competition-cooperation and subsequent hierarchical clustering showed aggregation of RBPs which could be interesting candidates for further investigation.

The stratification of intersections by length or overlap % and nearby binding sites by distance opened new insights on the interaction patterns of RBP pairs. Briefly, this new approach allowed us to accurately discern between small and big overlaps for intersections and short and long distances for nearby binding sites as shown in **Figure 2E and 2F**.

Discrepancies in the number of ranks when comparing the different metrics and consensus matrices denote the high complexity of RBP-RBP interactions. Whereas tendencies towards intersecting and nearby binding sites could be indicative of a generic competitive or cooperative profile respectively, these interactions are highly context-dependent. Hypothetically, RBPs cooperating for binding to a specific set of targets are not restricted to cooperative interactions exclusively throughout the whole transcriptome (Gardiner, Twiss, and Perrone-Bizzozero 2015; Gallardo-Dodd and Kutter 2024). In other words, if we assume RBP-RBP behaviour is context-dependent, then intersections and nearby sites might vary for different target RNA molecules and even different transcript regions, making it difficult to discern between generic competitive/cooperative profiles. This could explain why high

frequencies of big overlapping intersections can coexist with nearby binding sites and vice versa within the same RBP pair.

Nevertheless, the NMF clusters observed in **Figure 1.3** hint to general tendencies toward one type of interaction or the other. RBP pairs in cluster 1 manifest an accumulation of nearby sites of short distances (ranging from 0-10) which is indicative of close non-interfering binding. This particular feature of cluster 1 might suggest a potential cooperative cluster. Following a similar reasoning, cluster 3 displays an accumulation of intersections with small overlaps, mostly ranging from 0 to 10 %, Even if they intersect, the predominance of small overlaps suggest a recognition pattern of high dissimilarity between RBPs which makes competition for same binding sites relatively unlikely.

In contrast, cluster 2 is characterized by a distribution of intersections skewed towards high degree overlaps which is also consistent with the predominance of long intersections. In this case, we can say that this cluster is likely composed of potential competitive pairs. In addition, cluster 4 also exhibits a moderate tendency toward long intersections, even though the high-degree overlaps were not so prominent as in cluster 1, which could also be indicative of potential competitive profiles within this last NMF cluster.

The results concerning the NMF clusters, although inconclusive, point to the possibility of predicting cooperation and competition between RBPs. Nevertheless, the current analysis of segregated frequencies could be refined by attending to some criteria. For example, after the stratification of features, we could observe that some of the frequencies correlated across the RBP pairs - mostly those corresponding to the next or previous range of values-. In those cases, combining those ranges into one could simplify the analysis considerably. Moreover, after identifying the clusters, performing an exploratory analysis within each cluster to determine which RBPs tend to exhibit cooperative or competitive behaviors could provide valuable insights into the biological mechanisms underlying these interactions.

While these insights are promising, it is equally important to acknowledge the potential limitations inherent to both the eCLIP technique and the analysis pipeline, including the choice of cell lines, which may influence the interpretation of the observed RBP interactions. Despite mapping to targeted genes, eCLIP technique does not allow detection of the specific transcript variant where the binding occurs. Consequently, there is an undetermined fraction of reported RBP-RBP interactions which might be illusory events due to the presence of two RNA-binding sites in different transcripts**.**

In this work, cooperation and competition between RBPs have been exclusively addressed for HepG2, a cell line derived from cancer cells. ENCODE database currently accounts only for two cancer-derived cell lines with eCLIP data available, namely HepG2 and K562. While the current HepG2 model stands alone, the analysis of RBP-RBP interactions could be substantially enriched if other cell lines are considered, including not only K562, but also cell lines derived from normal cells still unavailable. The comparison of cancer against normal cells for the same RBP pairs could reveal shifts in interactions likely associated with cancer development and progression. In fact, there are numerous examples of RBP functionality being altered in cancerous cells (Lachiondo-Ortega et al. 2024; Velázquez-Cruz et al. 2021). To summarize, addressing differential binding across cell types could shed light on which post-transcriptional regulatory mechanisms involving RBP-RBP interactions are disrupted in cancer.

Materials & methods

Processing ENCODE eCLIP datasets

To access the eCLIP RNA-binding sites for the batch of 83 RBPs from HepG2 cell line, we downloaded the narrowPeak BED files for the two replicates available on the ENCODE database (Luo et al. 2020). We thus filtered eCLIP peaks with an accumulation of reads reported as statistically significant (i.e. $-\log_{10}$(p-value) > 1.30103). The identification of consensus peaks between the two biological replicates was performed by intersecting their pool of peaks and retaining only those peaks overlapping between replicates.

Computation of RBP-RBP interaction features

After preprocessing eCLIP peaks we computed all possible unique pairwise combinations of RBPs. Some RBPs were only included as RBP1 but not as RBP2 (e.g. AKAP1). To avoid any potential batch effect, prior to the computation of features, we switched RBP1 by RBP2 in all odd pairs so that each RBP was equally represented as RBP1 and RBP2. For each combination we identified the intersecting and nearby RNA binding sites by using bedtools 2.30 (Quinlan and Hall 2010). We considered binding sites of two RBPs as intersections when there exists a minimum overlap of at least one nucleotide. For the identification of nearby binding sites, we instead set a window of at most 100 nucleotides of distance

between two binding sites of both RBPs. The relative frequency of intersections or nearby binding sites was calculated dividing the number of events by the sum of binding sites of both RBPs. We computed the intersection overlap percentage by dividing the length of the intersection by the length of the binding site for each RBP. We calculated the median of binding site pairs values as a representative measure for the intersection length, overlap percentage, and nearby distance.

For each pair of RBPs, we stratified the reported set of intersecting and nearby binding sites. Each set of intersections was categorized based on the overlap percentage into ranges (0-100%) in increments of 10%. Similarly we measure the length of the overlapping sequence, dividing them into intervals ranging from 0 to 200 nucleotides with a step of 10 nts. For nearby binding sites (those within a certain proximity but not overlapping), we grouped them into distance ranges of 0-100 nucleotides using the same 10 nts step size.

Non negative matrix factorization (NMF)

NMF rank survey, generation of consensus matrices and the NMF for rank 4 were performed using the NMF R package (v 0.28) (Gaujoux and Seoighe 2010).

Chapter 2: The influence of potential post-translational modification sites on RBP behaviour.

Background

Post-translational modifications (PTMs) are covalent additions of functional groups to residues of a peptide chain. PTM deposition can trigger substantial changes in protein properties thus diversifying their functions. An increasing amount of work evidences the PTM dynamics as modulatory mechanisms of protein activity (Velázquez-Cruz et al. 2021; Geffen et al. 2023). Also RBPs are substrates of PTM deposition enzymes, representing a post-translational level of regulation of their activity. In fact, RBPs undergoing PTMs can drastically change their binding affinity for RNA molecules and other RBP partners (Velázquez-Cruz et al. 2021).

RBP-RNA interactions are complex events often involving more than one RBP leading to cooperative and competitive behaviors among them (Quattrone and Dassi 2019). An increasing amount of work indicates that cooperative and competitive interactions between RBPs modulate gene expression at post-transcriptional level and the consequent cell phenotypes (Quattrone and Dassi 2019; Khoroshkin et al. 2024). Addressing similar binding affinity or proximal disposition of binding sites of two RBPs is now feasible with the help of eCLIP datasets, providing a snapshot of the binding sites of an RBP in a cellular model (Van Nostrand, Pratt, et al. 2020b; Van Nostrand et al. 2016b; Khoroshkin et al. 2023).

However, whereas mapping coordinates of binding sites of a couple of RBPs could indicate potential interactions, whether these RBP-RBP interactions are the result of shared binding affinity for a target RNA or it is the PPI contributing to their formation remains to be answered. In this context, It is worth asking which are the molecular determinants shaping these interactions and how these alliances and handshakes are affected under specific diseases such as cancer (Dassi 2017). Taking this into account, in the present work, we thus focused on investigating the PTMs contribution to determining such interplay and the regulatory outcome.

Results

Potential PTM sites in RBPs preferentially accumulate at specific parts of the peptide and RNA-binding domains.

In RBPs, the most frequent PTMs are covalent additions of three functional groups: methyl, acetyl, and phosphate (**Supplementary Table 1.1A**). In this work, we analyzed annotated PTM sites in a set of 1350 human RBPs (see **Material & Methods**). To do so, we retrieved the location of PTM sites for each human RBP from UniProt (“UniProt: The Universal Protein Knowledgebase in 2023” 2023). Phosphorylation sites were found in 1005 human RBPs (74.74%), acetylation sites in 683 (50.22%), and methylation sites in 244 (18.07%) (**Figure 2.1B**). Other modifications are cumulatively found in less than 11% of all RBPs. Given their relative scarcity, these other PTMs are less likely to widely affect RBPs, and we thus selected the three most abundant PTMs for subsequent analyses.

Phosphorylation sites are widespread across RBPs, with a median of three sites per RBP, whereas the median amount of acetylation and methylation sites is zero (**Figure 2.1A-C, left**). PTMs are preferentially added to specific residues. We thus computed the PTM frequency for residues that can be modified and displayed the ratio of potential sites per residue (**Figure 1A-C, right**). To assess whether any PTM was enriched in RBPs, we computed a Fisher's test against the whole human proteome (20187 proteins). Interestingly, all but one tested PTMs were enriched (cysteine methyl ester, **Supplementary Table 1.1B**).

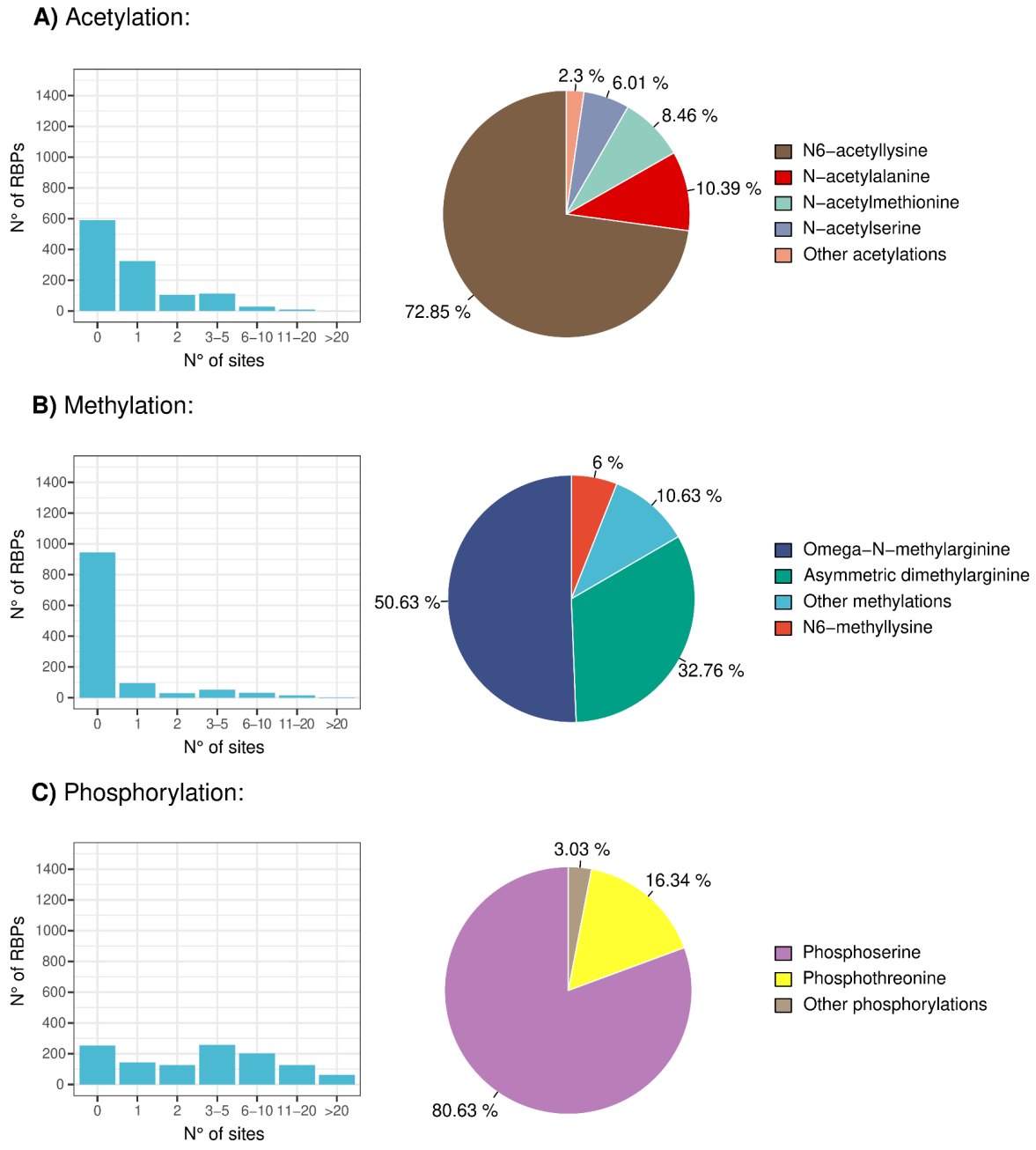


Figure 2.1: The distribution of the most abundant PTM types on human RBPs. *(A-C) show acetylation, methylation, and phosphorylation plots respectively. The bar plots in the*

left part indicate the abundance of potential PTM sites in RBPs. The x axis represents the number of sites per RBP and the y axis the number of RBPs with that specific number or range of PTM sites. The right pie charts represent the ratio of potential PTM sites per residue.

PTMs can determine protein function. As RNA-binding domains play a key role in defining the function of RBPs (Corley, Burns, and Yeo 2020; Gerstberger, Hafner, and Tuschl 2014), we could expect that PTMs in RBPs would be located within or close to such domains. In this context, we focused on the most frequent RBDs and regions in RBPs to carry out the analysis for the three main PTMs (**Figure 2.2A,B**). We thus evaluated the relative position and density of PTM sites within RBPs, across the full peptide length, the RBDs, and RNA-binding regions (**Figure 2.2C,D**). The distribution of PTM sites on RBPs appears to be non-random, particularly for acetylation and methylation. To evaluate an enrichment in PTM sites towards the N- or C-terminal of the protein, we splitted the peptide in two equal halves. We then grouped the PTMs according to their position relative to the center (right or left) and performed a t-test comparing nº of aa on the N-terminus versus C-terminus (N and C respectively). This revealed a significant accumulation of acetylation sites towards the N-terminal (FC(N/C) = 1.86 , p-value < 2.2e-16) (**Figure 2.2C, left**). This peak corresponds to the first amino acid, on which acetylation often stabilizes the peptide structure to ensure correct folding (Dandan Li et al. 2021; Lutz et al. 2017). Indeed, this enrichment holds also for non-RBP proteins. In contrast, methylation sites are enriched towards the C-terminal (FC(N/C) = 0.49, p-value < 2.2e-16) (**Figure 2.2C, center**), in line with the lesser structuration of the C-terminal, thus more accessible to methyltransferases than the N-terminal (J. Wang et al. 2023; Ivanov et al. 2007). In non-RBPs, however, methylation is not significantly enriched towards one or the other terminal (p = 0.26). Finally, phosphorylation sites are more homogeneously distributed. Still, we observe a modest enrichment towards the C-terminal (FC(N/C) = 0.93, p-value = 5.332e-05) (**Figure 2.2C, right**), as for the non-RBP proteins.

We then analyzed the PTM sites distribution in RNA-binding domains and regions (**Figure 2.2D**), defining a window of +/- 50 aminoacids around the region to include immediately upstream and downstream sites. Most PTM sites fall inside intrinsically disordered regions (IDRs). Interestingly, while methylation and phosphorylation sites are homogeneously distributed across the whole IDRs, acetylation sites accumulate at their beginning (**Figure 2.2D, left**).

Potential methylation sites tend to be close to RRM domains, being at a median distance of 26aas. Indeed, 82,87% of sites are downstream the domain and barely present inside RRMs (7.74 %). We can observe a similar trend also for the phosphorylation sites, with 75.44 % of them downstream or upstream of RRMs at a median distance of 22aas (**Figure 2.2D, right**). In the case of acetylation, the sites downstream and upstream of RRMs represent 56.78% of the total (**Figure 2.2D, middle**). Among all types of domains, the Zinc finger is the one for which PTMs are most frequently found around the RBD rather than inside it.

Finally, out of a total of 6861 reported PTM sites, only 2.32% were found within or nearby KH and DRBM domains and RNA binding regions, mostly due to the low abundance of these regions.

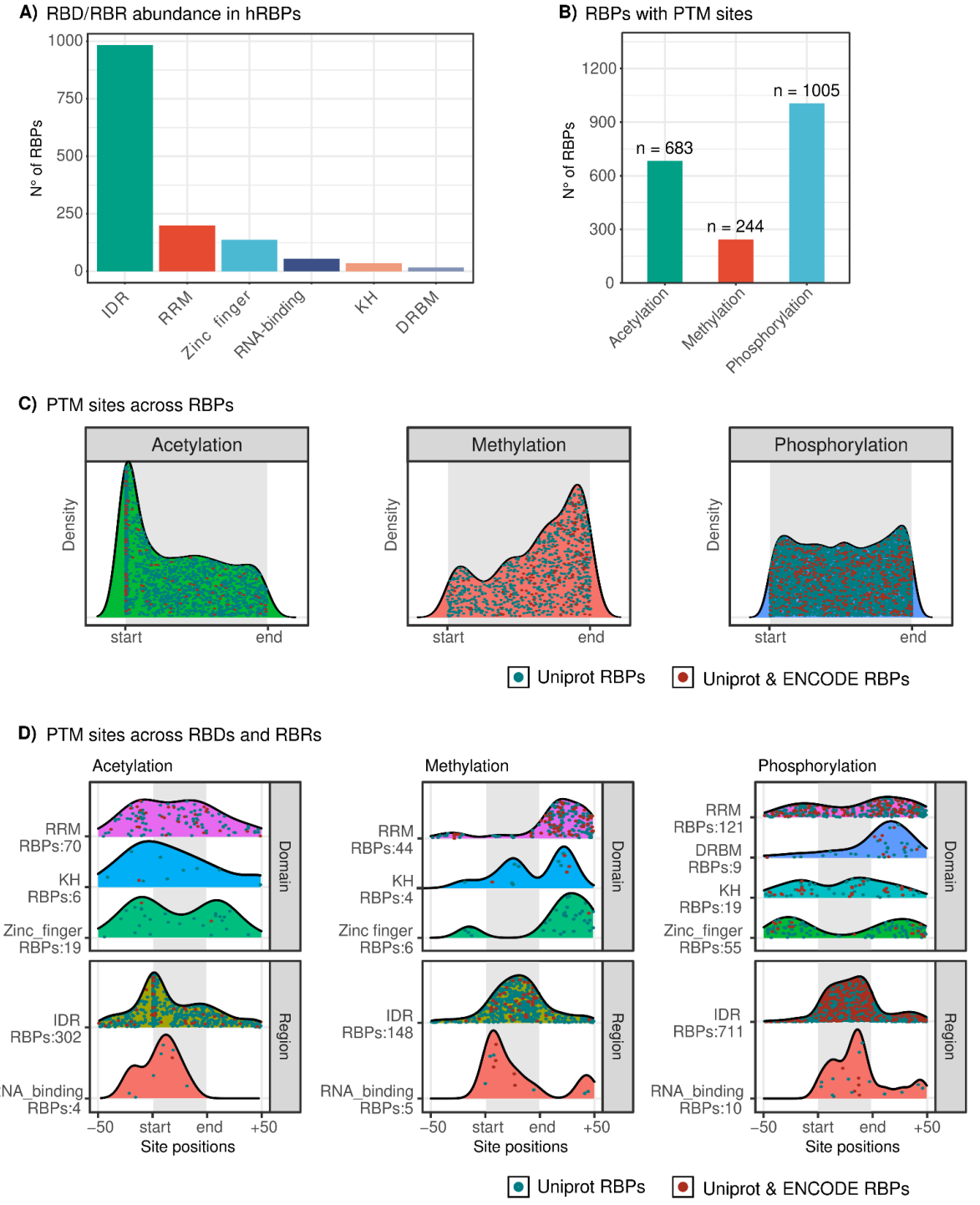


Figure 2.2: Distribution of potential PTM sites across the peptide length of RBPs. *(**A**) Bar plot showing the n° of RBPs containing different RBDs (or RBRs).The x axis represents the distinct RBDs or regions while the y axis indicates the number of RBPs. (**B**) Bar plot showing the n° of RBPs containing at least one potential PTM site among acetylation, methylation, and phosphorylation. (**C**) Density plots of the relative positions for potential PTM sites across the peptide chain. The x axis represents the peptide length scaled to 0-1 range from the N-terminal position (start) to the C-terminal position (end). Each point is a potential*

*PTM site colored according to the RBP source, i.e. red if the RBP is also found in the ENCODE dataset and blue otherwise (**D**) Density plots of the relative positions for potential PTM sites across the different RNA-binding domains and regions. The x axis represents the interval of the protein corresponding to RBDs and regions scaled to 0-1 range. The y axis represents the different domains and regions. Each point is a potential PTM site with the same color code as **C**.*

Acetylation sites in RBPs impact transcript region preferences whereas methylation sites increase the length of RNA-binding sites

Several PTMs that can modulate RBP-RNA interactions have been previously described (Velázquez-Cruz et al. 2021; Sternburg, Gruijs da Silva, and Dormann 2022). We thus sought to evaluate if the presence of potential PTM sites can have an impact on the RBP preference for its targets and which features could drive this behavior. To this end, we selected those RBPs with binding site-level data derived from the eCLIP technique in the HepG2 cell line from the ENCODE dataset (Luo et al. 2020), which contains a total of 83 such RBPs. Given the abundance of phosphorylation sites on RBPs we considered only acetylation and methylation for further analysis.

To test if PTM sites can make RBPs more interactive with RNA, we splitted the RBPs in two groups according to the presence/absence of potential methylation and acetylation sites and performed a Wilcoxon test. However, as shown in **Figure 2.3A,** the number of binding sites is not significantly different (acetylation: FC(no-site/site) = 0.7292, p-value = 0.1528, methylation: FC(no-site/site) = 2.0741, p-value = 0.0939) . If instead we compare the binding site length between groups, the RBPs with methylation sites showed significantly longer binding sites than those without (FC(no-site/site) = 2.0741, p-value = 0.012) (**Figure 2.3B**).

We thus sought to test whether the sequence and structure content of the binding sites could be associated with differential presence of PTM sites. In particular, we assessed the association of potential PTM sites in RBPs with the recognition of secondary structure elements in their binding sites. To this end, we computed the relative frequency of each possible structural element (i.e. stem, internal loop, hairpin loop and multiloop) inside the binding sites and obtained the median for each element. Finally, we performed a Wilcoxon test comparing the two groups of RBPs (presence/absence of potential PTM sites) using

these median values (**Figure 2.3C**). The results showed no significant difference between groups for any of the structural elements.

Then, to understand if PTM sites could affect the RBP preference for transcript regions, we computed the relative binding frequency of each RBP on the different regions. We observed that 70 RBPs (80,72%) fell in two groups, those that preferentially bind intronic regions and those most frequently binding the CDS (**Figure 2.3D**). Only 8 RBPs in the dataset had as first preference the 3'UTR region and just one the 5'UTR.

We then split the RBPs in two groups, those with at least one potential PTM site and those with none. We clustered the RBPs within each group according to their relative binding frequency in each transcript region (**Figure 2.3E-G**), and assessed a possible effect of PTMs on the number of binding sites at each transcript region with a t-test. Even though no significant difference was found for methylation, we observe significant differences between RBPs with potential acetylation sites vs those with none. In particular, RBPs with acetylation sites have significantly less binding sites in the 3'UTR and more in unclassified exons (fold-change = 1.37x, p-values: 0.017 and 0.019 respectively, **Figure 2.1F**).

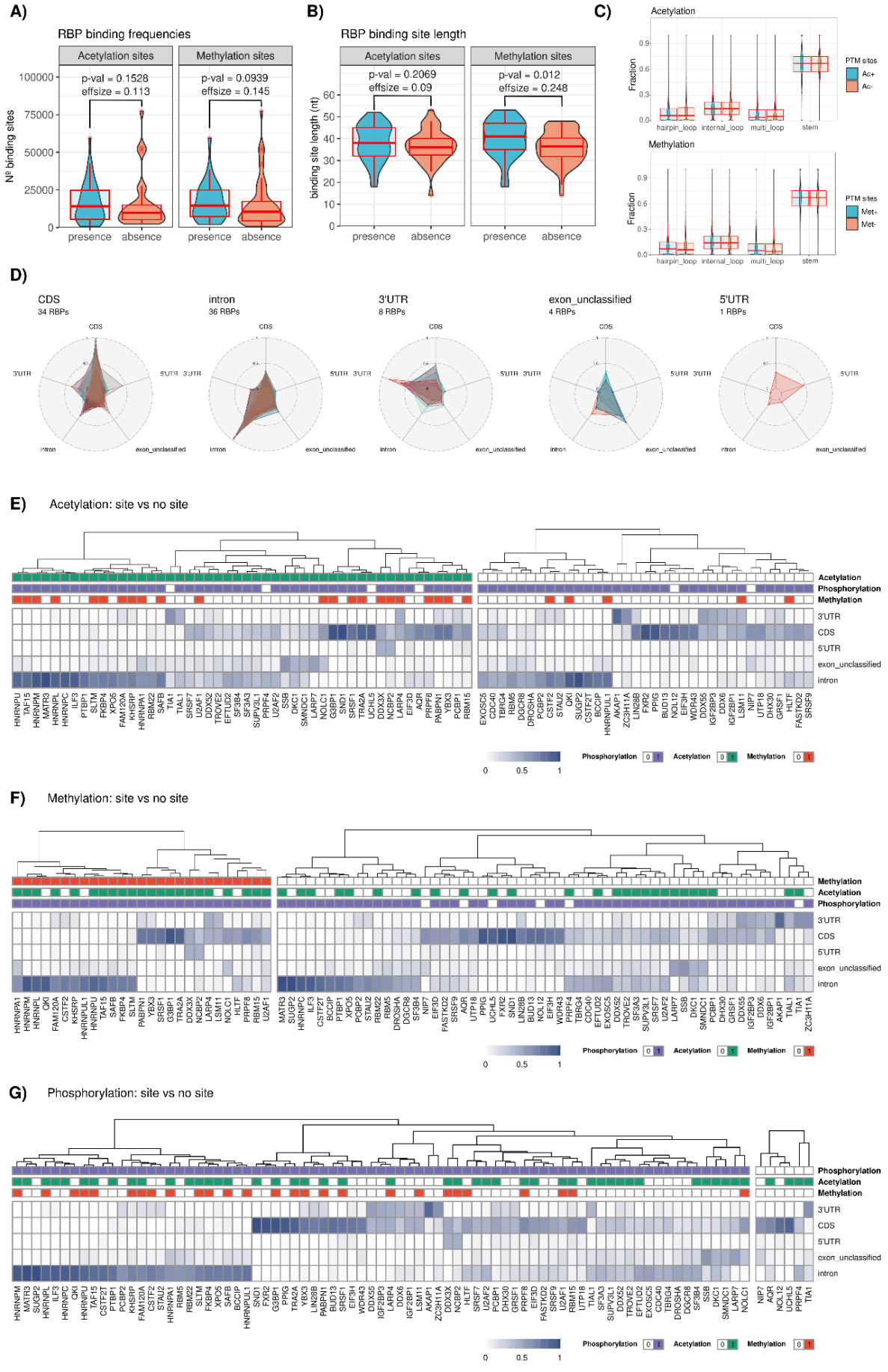

A)
RBP binding frequencies
Acetylation sites
Methylation sites
p-val = 0.1528
effsize = 0.113
p-val = 0.0939
effsize = 0.145
Nº binding sites
presence
absence
B)
RBP binding site length
p-val = 0.2069
effsize = 0.09
p-val = 0.012
effsize = 0.248
binding site length (nt)
C)
Acetylation
Methylation
Fraction
PTM sites
Ac+
Ac-
Met+
Met-
hairpin_loop
internal_loop
multi_loop
stem
D)
CDS
34 RBPs
intron
36 RBPs
3'UTR
8 RBPs
exon_unclassified
4 RBPs
5'UTR
1 RBPs
E) Acetylation: site vs no site
F) Methylation: site vs no site
G) Phosphorylation: site vs no site
Phosphorylation
Acetylation
Methylation
3'UTR
CDS
5'UTR
exon_unclassified
intron

Figure 2.3: Comparison of RBP-RNA interactions in presence/absence of potential PTM sites. *(**A**) Violin box plots of the number of RNA binding sites for RBPs. Comparison between groups of RBPs in presence (blue) vs absence (red) of potential acetylation (left) and methylation (right) sites. (**B**) Violin box plots of the binding site length for RBPs. Same conditions that A. (**C**) Violin box plots for the fractions of the different RNA secondary structures found on the binding sites. Same conditions that A and B. (**D**) Radar plots displaying the RBP binding profiles grouped according to their preferential transcript regions. (**E-G**) Heatmap of hierarchical clustering of RBPs based on their relative binding frequencies on each transcript region. The RBPs on the x axis are separated according to the presence of acetylation, methylation and phosphorylation in E, F and G heatmaps respectively.*

PTMs in RBPs are differentially represented in molecular functions and subcellular components

PTMs can be determinant for protein transport within the cell, with direct consequences on RNA metabolism (Velázquez-Cruz et al. 2021; Long et al. 2019; Thapar 2015). Therefore, we asked ourselves if the presence of potential PTM sites on RBPs could be associated with their localization to specific subcellular compartments. To answer this question, we first extracted the subcellular location annotations from the Human Protein Atlas (HPA) (Thul et al. 2017), with 1219/1350 RBPs found in the database (90.2%). We then performed Fisher's exact tests comparing the fraction of PTM site-containing RBPs in each subcellular location against the proteome-wide fraction of all RBPs (**Figure 2.4C**). We only detected a significant enrichment in the case of phosphorylation and acetylation sites on RBPs located in nuclear speckles (7.63% of RBPs, p-values: 0.0011 and 0.0479). The test also revealed significant depletions of PTMs in three subcellular locations. In particular, RBPs in vesicles and mitochondria are significantly depleted in phosphorylation (6.48% and 10.17% of RBPs; p-values: 4.8692e-03 and 1.6486e-10) whereas mitochondria and nucleoli are depleted in methylation (10.17% and 18.54% of RBPs; p-value: 0.0001 and 0.0106). Only nucleoli are significantly depleted in acetylation (18.54% of RBPs, p-value = 0.0064).

In addition to subcellular location, another important aspect of RBPs where PTMs can play a key role is their function (Long et al. 2019; Daitoku, Sakamaki, and Fukamizu 2011). To examine possible associations between PTM sites and RBP functions we extracted the GO annotations for each RBP from the Uniprot database. For the following analysis we considered GO terms from the biological process (BP) and molecular function (MF)

branches. We performed Fisher's exact test for each GO term comparing the abundance of PTMs in those RBPs annotated with a particular term versus the abundance in all RBPs. All the annotations with BH-adjusted p-value less than 0.01 were considered significantly enriched or depleted in PTMs (**Figure 2.4A**). Among the GO terms that were significant in any of the PTMs tested, we identified closely related terms that could be classified in four subgroups: mRNA binding, protein aggregation, protein folding and translation (**Supplementary table 1.2**).

Then, to study how RBPs can group based on their functions, we applied UMAP to the RBP dataset, retaining those annotations with at least three occurrences. We marked RBPs with the most frequent molecular functions in the dataset, and performed a k-nearest neighbors classification using the coordinates of the UMAP components to find the 15 closest neighbors of each RBP. (**Figure 2.4B**). Interestingly, the RBP clusters we obtained recapitulated the functional subgroups previously identified by GO analysis.

Finally, we asked ourselves whether the RBP clusters related to different functions could rely on protein-protein interactions (PPIs) of the same type. PTMs can modulate this type of interactions by, for instance, inducing conformational changes in interacting proteins, thus modulating the occurrence of PPIs (Egbert et al. 2023; Thapar 2015). We thus explored the PPI interaction network between RBPs from the same functional subgroups previously identified (**Figure 2.4A**) by extracting experimentally validated PPI for these proteins from STRING (Szklarczyk et al. 2019) (**Figure 2.5A-D**). For these two clusters, no significant preference for a specific PTM was observable, with acetylation being however considerably more frequent than other modifications in the translation cluster.

Figure 2.4: PTM enrichment analysis across multiple GO terms and subcellular locations. A) *Heatmap of log(p-value) for each Fisher's exact test performed on GO terms. Red and blue indicate significant PTM enrichment and depletion respectively. (**B**) UMAP projection of RBPs from Uniprot database, colored according to the most predominant MFs. RBPs without any of the MF annotations plotted were labeled as "Others". (**C**) Stacked bar plots for each subcellular location of RBPs in presence vs absence of PTM sites.*

Having explored how RBP biology could be influenced by the presence of potential PTM sites, we now focus our attention on the relation between PTM sites and RBP-RBP interactions. We extracted experimentally validated PPIs for all RBPs from STRING, amounting to 1347 RBPs and 27166 interactions and obtained node degrees for each RBP in the network (Szklarczyk et al. 2019). Three RBPs (SNHG28, HSPA7 and ZRSR2P1) did not have any interactions.

We thus compared the distribution of annotated PPIs per RBP between the groups of PTM RBPs and the non-PTM RBPs. This revealed that RBPs with acetylation sites had significantly higher numbers of PPIs than the non-acetylation sites RBPs (FC(acetylation/not acetylation) = 1.275x, U Mann Whitney p-value = 2.932e-06). No significant difference was instead observed for methylation (FC(methylation/ not methylation) = 0.828x, U Mann Whitney p-value = 0.9971).

This analysis suggested a possible effect of acetylation sites on RBP PPIs at a global scale. We thus focused again on the four functional subgroups of RBPs. First, we compared PPIs within the subsets of RBPs versus the full RBP network. This showed that the translation subnetwork had a significantly higher PPIs/RBP ratio than globally (FC = 1.3476, U Mann Whitney p-value = 2.09e-08), whereas mRNA binding, protein aggregation and protein folding subgroups showed significantly less PPIs/RBP (FC: 0.2203, 0.0533 and 0.3379; p-values: 1.3199e-15, 1.7394e-25 and 0.022 respectively). (**Supplementary table 1.3A**).

Next, we compared PTM versus non-PTM protein sets within subgroups. This showed that RBPs without acetylation sites have significantly less PPIs than RBPs having them within the mRNA binding and protein folding subnetworks (FC: 2.2124 and 2.2621; U Mann Whitney p-values: 2.6024e-10 and 0.0011) . In the case of methylation, RBPs without sites have significantly less PPIs than RBPs with them within mRNA binding, protein aggregation and protein folding subnetworks (FC: 1.2492, 2.7541, 1.7318; p-values: 0.0126, 0.0002 and 0.0059) , but significantly more in the case of the translation subgroup (FC = 0.5271, p-value = 0.0004) (**Supplementary table 1.3A**).

When we study the effect of PTMs at the whole proteome level rather than only RBPs, the presence of acetylation and methylation sites implies a significantly higher number of PPIs (FC: 2.7425 and 2.515; p.value < 2e-16 in both cases) (**Supplementary table 1.3B**). Even though acetylation sites seem to promote PPIs, as observed for RBPs, the effect of

methylation sites in RBPs appears to be absent. When RBPs with potential acetylation and methylation sites were compared against the whole set of proteins with potential sites we observed that overall RBPs have significantly less PPIs than other proteins (p-value < 2.2e-16 in both cases, FC(mean(PPIs per RBP)/ mean(PPIs per protein)): 0.1368 and 0.0932 respectively).

Figure 2.5: RBP functional subnetworks obtained from STRING database based on GO terms associations. *(**A-D**) Functional subnetworks of RBPs involved in protein folding, mRNA binding, protein translation, and protein aggregation respectively. Blue red and purple RBPs carry acetylation, methylation or both potential sites respectively. Only PPIs experimentally validated are displayed.*

The presence of potential PTM sites on RBPs can discriminate profiles of RBP-RBP interaction.

The presence of PTM sites appears to be related with the number of PPIs made by the containing RBPs. We therefore asked ourselves whether these sites could also affect post-transcriptional RBP-RBP interactions, i.e. the potentially cooperative or competitive binding and regulation of the same target RNAs. To this end, we analyzed the intrinsic features of those interactions, by exploiting RBP binding sites information from the 83 RBPs with ENCODE eCLIP data for the HepG2 cell line as described in the **Material & methods** section from Chapter 1 (Luo et al. 2020). We analyzed all pairwise combinations of these RBPs, obtaining a total of 3403 pairs, and computed a set of more than one thousand features for each pair.

Given the high number of computed features, including several describing the relative distance of binding sites of the RBPs in the pair, we applied dimensionality reduction techniques to discern possible RBP pair clusters. As can be observed on the tSNE projection in **Figure 2.6A**, the presence of methylation and acetylation sites in one or both RBPs of the pair appear to be a driving factor of pair clustering.

Among these features, we focused on the frequency of intersecting binding sites between the two RBPs (intersection) and of proximal but non-overlapping binding sites (nearby). A comparison between groups of RBP pairs (with none, one and both containing potential PTM sites) showed a significant decrease in nearby peaks and intersection frequencies when acetylation and methylation sites are present in one or both RBPs (**Figure 2.7C**). In particular, pairs with acetylation sites in one or both RBPs showed significantly less intersections than pairs where none of the RBPs had acetylation sites (FC: 1.17x and 1.38x; p-values: 5.3e-07 and < 2e-16 respectively). The comparison between one and both was also significant (FC = 1.18x, p-value = 3.0e-07). Similarly, pairs with one and both RBPs containing methylation sites were also significantly depleted in intersections with respect to non-methylated RBP pairs (FC: 1.2x and 1.32x; p-values: 2.3e-9 and 1.3e-07 respectively). Finally, the groups with both or one RBPs containing methylation sites did not significantly differ for the intersection frequency.

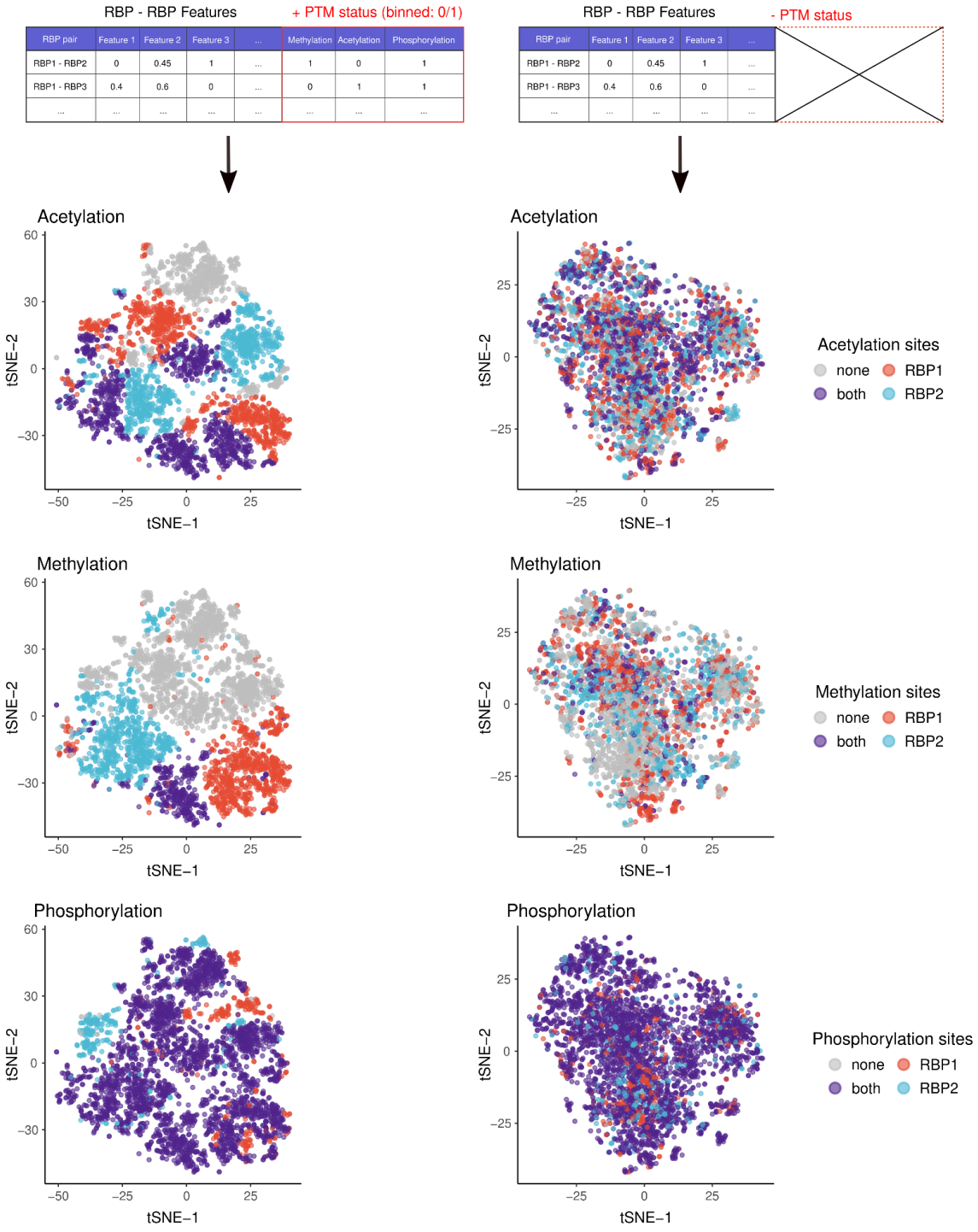


Figure 2.6: Unsupervised clustering of RBP-RBP pairs in presence and absence of PTM features. *tSNE projections of RBP pairs including (left column) and excluding (right column) 0/1 binned PTM information, i.e. presence/absence of potential sites for acetylation, methylation and phosphorylation. The PTM status for each pair is indicated with different colors, gray (none), red (RBP1), blue (RBP2) and purple (both).*

To test how the different features associate with the presence of PTM sites, we systematically computed U Mann Whitney tests. To perform the comparisons we splitted the RBP pairs in three subsets according to the presence of PTM sites in one, both or none of the two RBPs. Interestingly, we could observe significant differences for several features when comparing these three groups for acetylation and methylation sites. (**Figure 2.7A,B**).

Among these features, we focused on the frequency of intersecting binding sites between the two RBPs (intersection) and of proximal but non-overlapping binding sites (nearby). A comparison between groups of RBP pairs (with none, one and both containing potential PTM sites) showed a significant decrease in nearby peaks and intersection frequencies when acetylation and methylation sites are present in one or both RBPs (**Figure 2.7C**). In particular, pairs with acetylation sites in one or both RBPs showed significantly less intersections than pairs where none of the RBPs had acetylation sites (FC: 1.17x and 1.38x; p-values: 5.3e-07 and < 2e-16 respectively). The comparison between one and both was also significant (FC = 1.18x, p-value = 3.0e-07). Similarly, pairs with one and both RBPs containing methylation sites were also significantly depleted in intersections with respect to non-methylated RBP pairs (FC: 1.2x and 1.32x; p-values: 2.3e-9 and 1.3e-07 respectively). Finally, the groups with both or one RBPs containing methylation sites did not significantly differ for the intersection frequency.

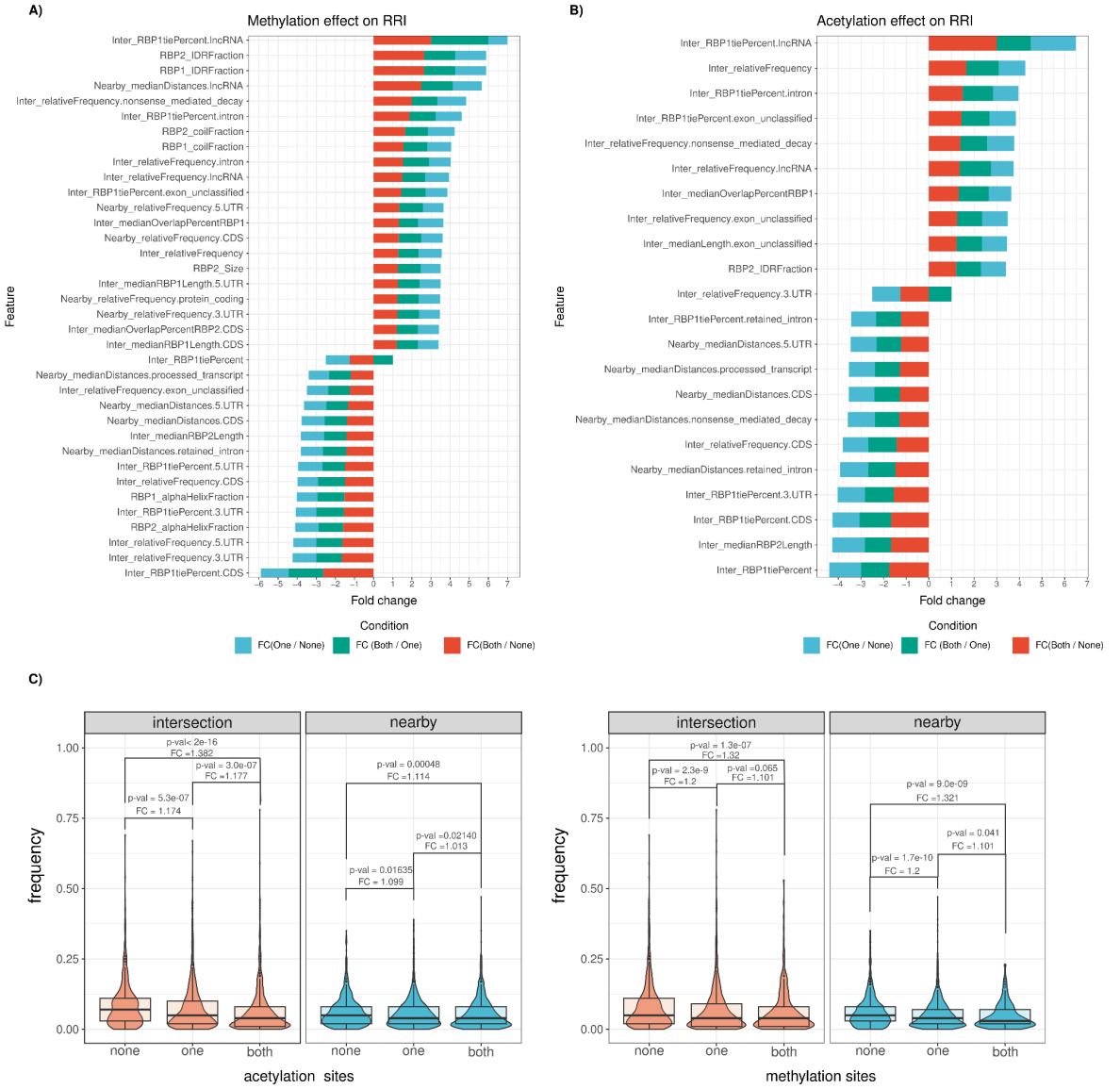


Figure 2.7: PTM effect on RBP-RBP interactive features. *(**A,B**) Stacked bar plots representing the fold change between different PTM status (A, B corresponding to methylation and acetylation respectively) observed for each RBP-RBP feature. Red for both versus none comparison, green for both versus one and blue one versus none. Features under a fold change of 1.2 or p-value greater than 0.05 are not shown. (**C**) Comparison of intersecting (left) and nearby (right) binding sites frequencies between RBP pairs with none, one or both RBPs carrying potential PTM sites. P-values and effect sizes reported correspond to the Wilcoxon test for independent samples.*

The tSNE projections shown in **Figure 2.6** revealed that the presence of PTM sites can potentially cluster the set of RBP pairs in several subgroups. Additionally, we have seen that the RBP-RBP interaction features are significantly affected by the PTM sites presence (**Figure 2.7A,B**). We thus asked ourselves whether these features could be used to predict if any two RBPs could interact through binding sites on their target RNAs.

To answer this question, we used XGBoost (T. Chen and Guestrin 2016) and implemented a supervised learning approach predicting the frequencies of intersecting and nearby binding sites and the presence of PTM sites. We trained XGBoost in regressor mode to predict intersection and nearby relative frequencies, and in classifier mode to predict the PTM sites presence for the RBP pair. We carried out a grid search to determine the best hyperparameter values, which improved the performance of approximately 5% for the different evaluation metrics in both the regressor and classifier modes (**Supplementary table 1.4**).

The regressor model performance was measured using the residual mean squared error (RMSE) at each boosting iteration. The final score for the intersection and nearby binding site frequencies was 0.063 and 0.05 respectively (**Supplementary table 1.4**). We thus ranked the features according to their information gain. The median distance between nearby binding sites resulted to be the most informative feature, followed by the median length of the RNA-binding sites. It is worth mentioning the motif similarity score appears to be an important feature to predict intersection frequencies but not for nearby binding sites (**Figure 2.8A,B**).

The classifier model had a better overall accuracy when methylation sites are predicted with respect to acetylation (accuracy: 0.912 and 0.82 respectively). However the acetylation classifier predicted better pairs where both RBPs had PTM sites. Ranking features showed that the RBP co-expression across tissues and several structural features have the highest importance scores when predicting the acetylation sites (**Figure 2.8C**). The RBP size and IDR fraction were instead the most important for methylation sites (**Figure 2.8D**).

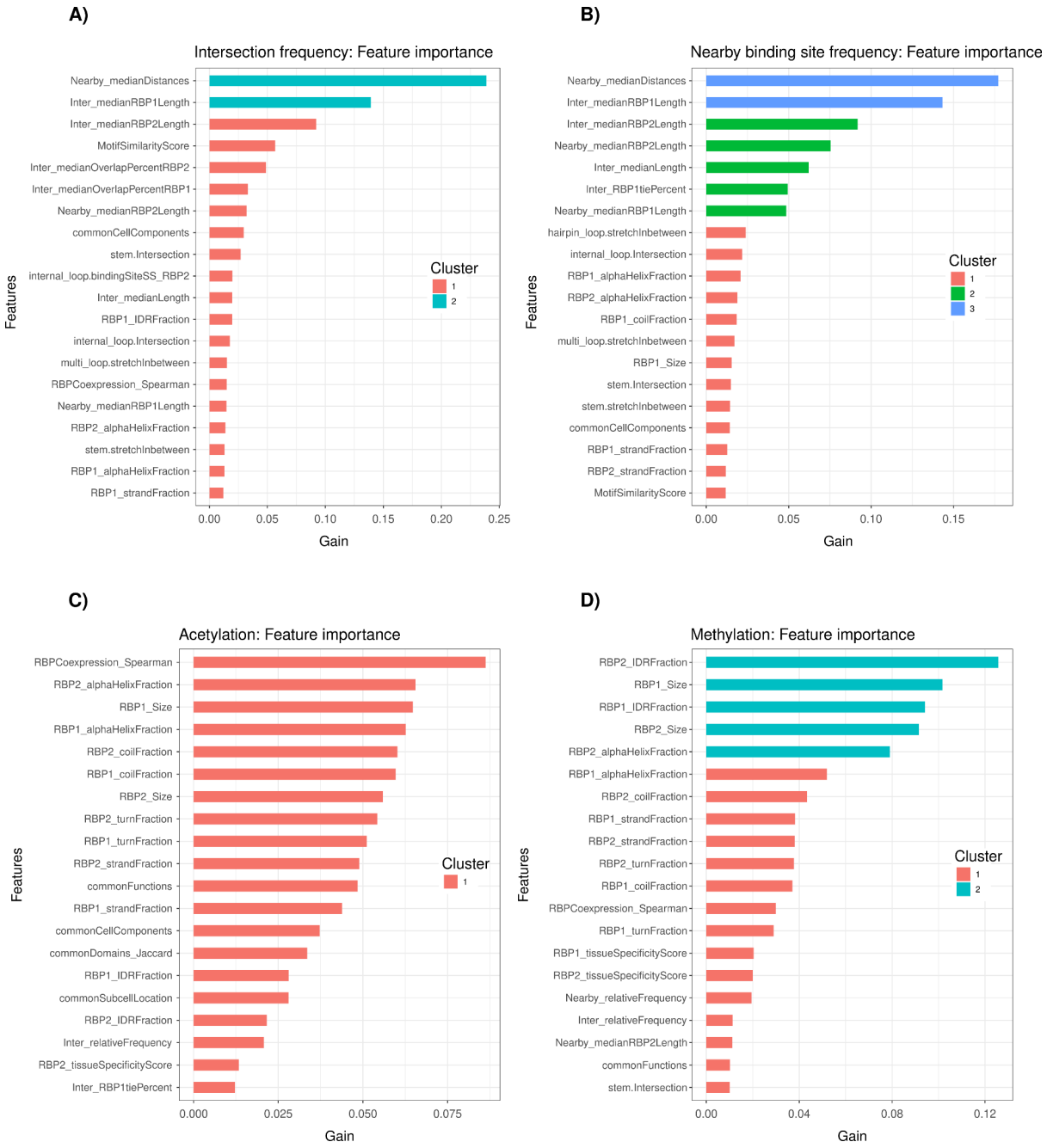


Figure 2.8: Evaluation of features contribution to XGBoost classification and regression tasks *(**A,B**) Top 20 XGBoost regressor features used to predict the intersection and nearby binding site frequencies ordered by level of importance in terms of information gain (**C,D**) Top 20 XGBoost classifier features for prediction of PTM status ordered by information gain.*

RBP-mediated post-transcriptional regulation of PTM enzymes.

RBPs can bind and regulate the transcripts coding for PTM deposition and erasing enzymes. As the correct balance between PTM writers and erasers can determine protein function and its alteration lead to disease onset, their post-transcriptional regulation could be crucial to maintain homeostasis (Sternburg, Gruijs da Silva, and Dormann 2022; Geffen et al. 2023). We thus hypothesized the RBP participation in post-transcriptional regulation of PTM enzymes, subsequently modulating PTM deposition activity. In this context, our findings suggest the existence of regulatory feedback loops where RBPs can tune the PTM deposition rates and thus their own activity and interactions.

The list of enzymes involved in acetylation and methylation was extracted from the HGNC website (Seal et al. 2023). The resulting list was composed of 80 acetylation enzymes (55 acetyltransferases and 25 deacetylases) and 108 methylation enzymes (78 methyltransferases and 30 demethylases). We then intersected the genomic coordinates of those genes with RNA-binding sites from the ENCODE eCLIP datasets to identify which RBPs were significantly enriched in targets for each enzyme class. According to the nº of PTM deposition genes targeted, three subgroups of enriched RBPs were defined, i.e. high, medium and low targeting frequency (**Figure 2.9A**).

First of all, we quantified the number of RBPs post-transcriptionally targeting PTM enzymes. Our results revealed 45 acetyltransferases and 19 deacetylases out of 80 acetylation enzymes targeted by at least one of the 103 RBP in the list of HepG2 eCLIP experiments. From the pool of 108 methylation enzymes, 60 targeted methyltransferases and 22 demethylases were reported. The higher degree of targeting toward writer enzymes compared to erasers that we observed could be attributed to the underrepresentation of erasers within the whole pool of enzymes. From an RBP-centric approach, we identified 90 RBPs out of 103 having RNA-binding sites in at least one listed PTM gene. In particular, 81.55 % of RBPs targeted acetyltransferase genes, 73.79 % deacetylases, 84.47 % methyltransferases and 72.82 % demethylases.

To evaluate potential enrichment in each of the four enzyme categories, we performed a one-sided Fisher's exact test with a level of significance of 0.05 for each of the RBPs using the ratio of the whole set of targeted genes divided by the eCLIP background against the ratio of targets within each enzyme type:

$$\frac{nº\ of\ targets}{eCLIP\ background}\ v.\ s.\ \frac{nº\ of\ targeted\ PTM\ enzymes}{total\ nº\ of\ enzymes}$$

The statistical test revealed 60 RBPs significantly enriched in at least one of the enzyme groups. Out of those 60 RBPs, 31 were enriched in acetyltransferases genes, 20 in deacetylases, 42 in methyltransferases and 47 in demethylases. The 70% of enriched RBPs were enriched in more than one category. In particular, 20 RBPs were enriched in both acetylation and methylation writers whereas 18 were enriched in erasers. Finally, a subgroup of 11 RBPs was enriched in all enzyme classes.

As part of a hypothetical regulatory loop, RBPs carrying potential PTM sites were expected to be more involved in PTR of PTM deposition enzymes compared with non-PTM-site RBPs. However, no significant differences were found between PTM-sites and non-PTM-sites RBPs in terms of the number of PTM genes targeted suggesting an equal contribution to the regulation of PTM deposition.

In line with the previous results, differences in RBP-RBP interactions involving PTM-enzyme enriched RBPs were also evaluated. To this end, we considered again our set of features computed for each RBP pair. The previous dataset of RBP pairs was splitted in pairs with at least one of them enriched in PTM enzymes versus the pairs with none of the RBPs enriched. As shown in **figure 2.10A,B**, the pairs of RBPs with target-enrichment in PTM genes presented significantly less intersection and proximity events when compared against not enriched pairs, suggesting a less interactive profile.

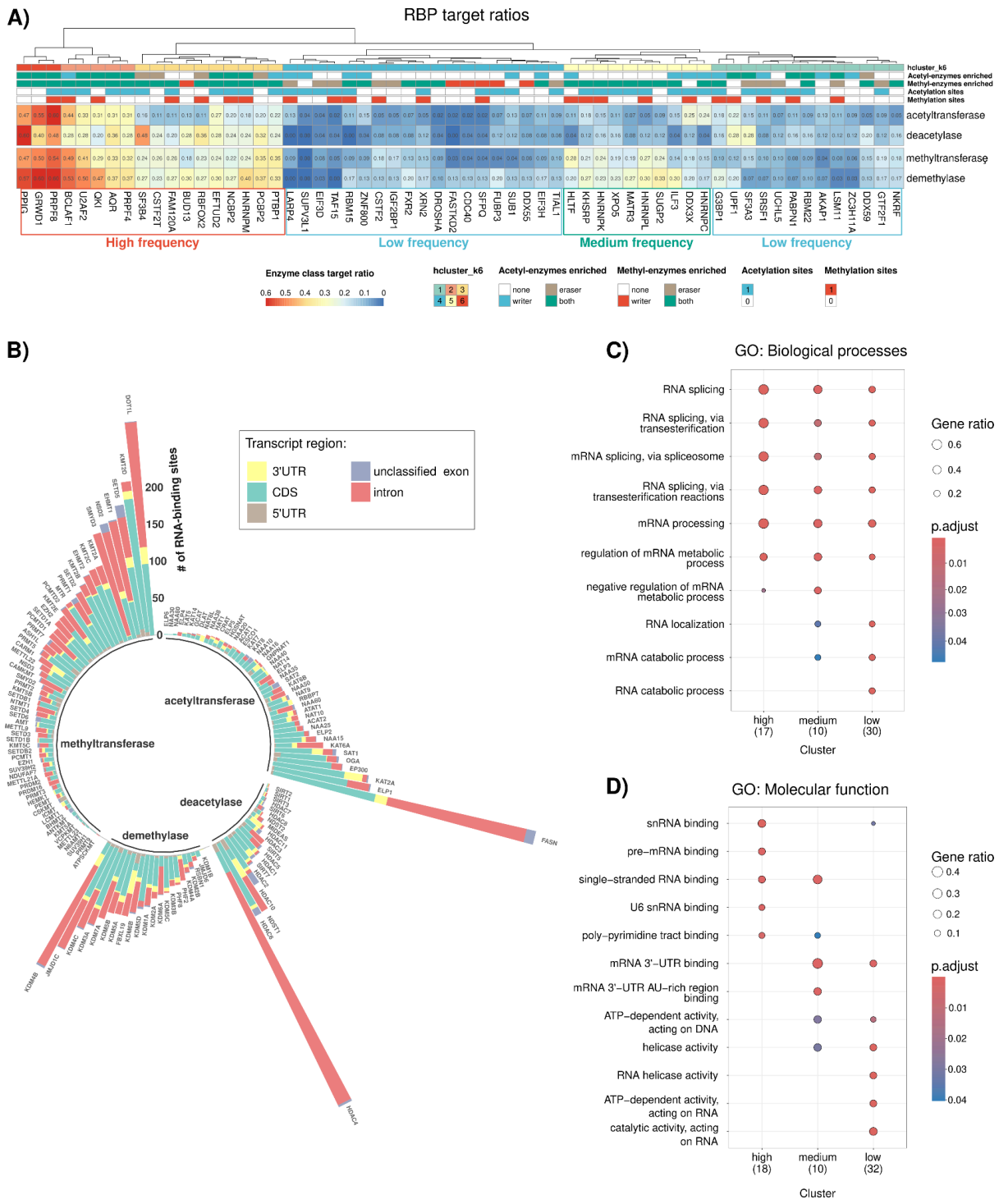


***Figure 2.9: RBP participation in post-transcriptional regulation of PTM deposition enzymes**: (**A**) Heatmap of RBPs enriched in PTM enzyme targets. Each cell value represents the ratio of targeted enzymes from a specific class for a single RBP. Hierarchical clustering was applied to RBPs revealing three main groups according to their ratios, i.e. low medium and high frequency clusters. (**B**) Barplot showing the frequency of RBP-RNA interactions throughout the transcripts encoding PTM deposition enzymes. Each bar represents a PTM deposition gene and these are grouped by their respective enzymatic*

*types. Each color within the bar represents a different transcript region. (**C**, **D**) Dotplots for enriched biological processes and molecular functions respectively. The plots show levels of significance for most represented GO terms among the PTM-enzyme enriched RBPs. BH adjusted p-values are represented by a blue to red color gradient from low to high significance level respectively and the size represents the ratio of RBPs annotated with the GO term.*

In order to understand the RBP-enzyme interplay, we performed a GO enrichment analysis over the enriched RBPs. The analysis of GO terms revealed an overall enrichment in the RNA splicing and mRNA processing functions for these RBPs (**Figure 2.9C, D**). We thus evaluated whether those RBPs targeting more PTM enzymes were more related to the splicing function than those targeting less enzymes. The comparison between these three groups of RBPs showed significantly higher odds ratios of splicing-related GO terms in the RBPs with high targeting frequency versus medium and low ones (p-value = 0.016 and 7.3 e-06, FC = 1.38x and 3.87x respectively) (**Figure 2.10C,D**). Interestingly, the current results not only suggest that RBP regulation of PTM deposition genes is mediated mostly via alternative splicing, but also that a greater participation in the PTR of those genes is associated with RNA splicing functions.

If we map the interactions between RBPs and PTM enzymes we observe some enzymes standing out among the pool of PTM enzymes displayed in the heatmap (**Figure 2.9A**). In particular the fatty acid synthase I (FASN) is the acetyltransferase with the highest number of RBP binding sites, mostly mapping on intronic regions (**Figure 2.9B**). The overexpression of FASN is correlated with tumor progression in breast cancer (Alwarawrah et al. 2016). The most abundant RNA-binding sites mapping within FASN belong to the RBPs AQR, PRPF8 and EFTUD3 which have been recently reported as common interactors of the spliceosomal protein SNRNP200 (L. Street et al. 2023). In the same line, further analysis revealed that FASN is significantly enriched in intersections of AQR and PRPF8 RNA-binding sites compared to the pool of intersections in the whole transcriptome, suggesting a likely interplay in the PTR of this enzyme.

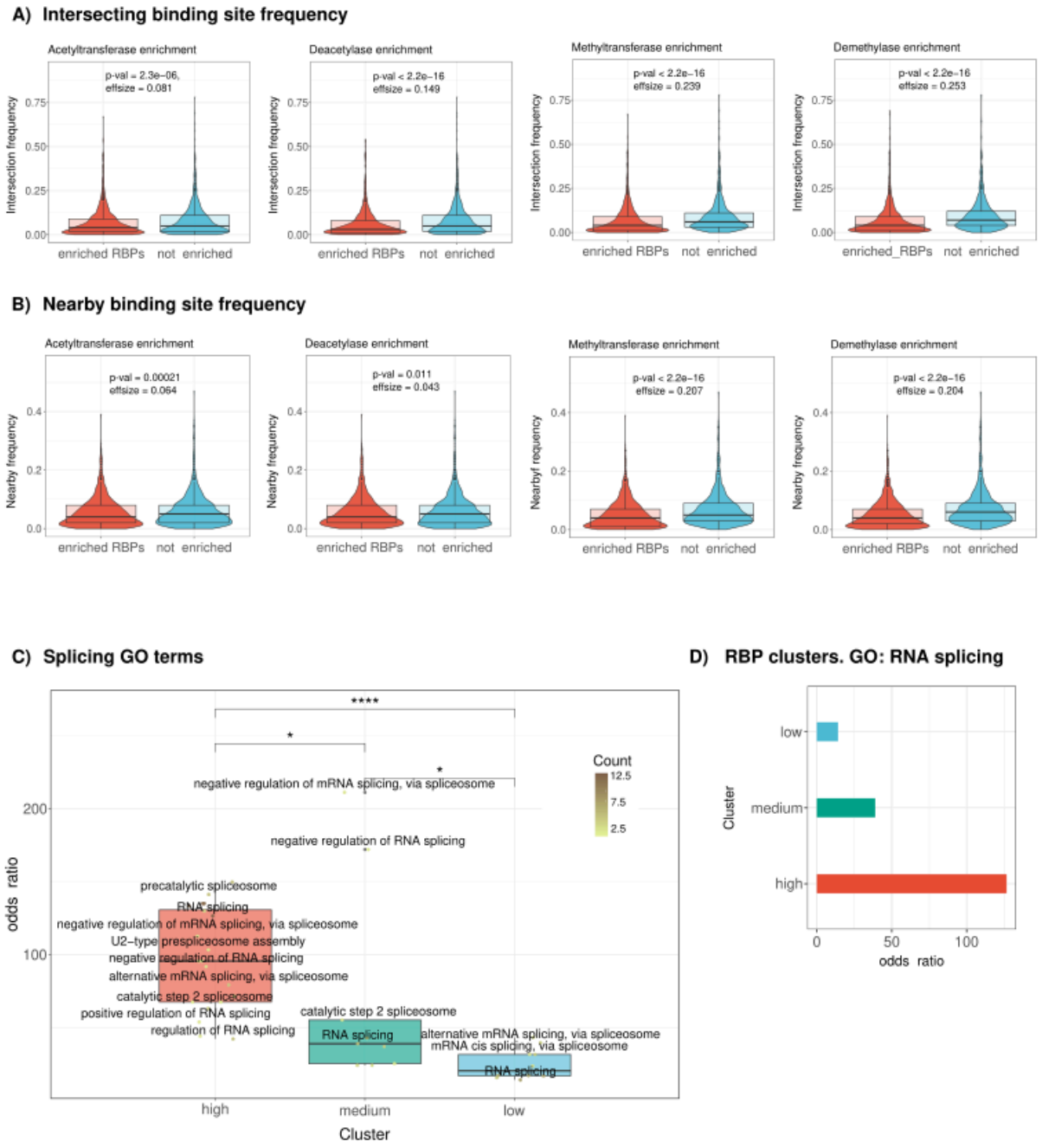


Figure 2.10: Features of RBPs enriched in PTM enzyme targets *(**A-B**) Boxplot: Intersection and nearby binding site frequencies (upper and lower rows respectively) of RBP pairs including at least one RBP significantly enriched in one of the PTM deposition gene lists (i.e. acetyltransferases, deacetylases, methyltransferases and demethylases) versus not enriched RBP pairs. Each boxplot panel split the data into enriched and non enriched for a specific enzyme class. The p-values and effect sizes displayed correspond to the Wilcoxon test for independent samples. (**C**) OR boxplots. Comparison of splicing related GO terms between the three RBP clusters identified on previous heatmap. (**D**) Barplots showing the difference in OR for the RNA splicing term in the three clusters.*

Discussion

An intriguing aspect of PTR that remains poorly studied is the role of PTMs in shaping the RBP-RBP interactions. Our study of PTMs provides insights on the influence of these yet unappreciated factor in cooperative and competitive behaviours of RBPs. PTM regulation of protein function indeed represents a potential additional layer of complexity in regulation of gene expression. While our findings shed light on some aspects of RBP dynamics, they also highlight the necessity of further exploring PTMs as modulators of these interactions.

Acetylation, methylation and phosphorylation sites are the most widespread PTMs among RBPs. More precisely, the serine/threonine phosphorylation and the lysine acetylation are the most abundant and conserved PTMs, not only in RBPs, but also in the whole proteome in eukaryotic cells (Geffen et al. 2023). Furthermore, the three PTMs exhibit defined patterns of distributions throughout the polypeptide chain, suggesting location is associated with their specific roles as modulators of the RBP activity.

From a functional point of view, the accumulation of acetylation sites towards the N-terminal could be explained by the protein stabilization function of this PTM. In fact, acetylation of the N-terminal amino acid serves as a stabilizer of the peptide structure and ensures correct folding (Arnesen 2011; Velázquez-Cruz et al. 2021). Even more interesting is the fact that the C-terminus is often involved in PPIs. Therefore PTMs across the C-terminus, and in particular methylations, could have a role as modulators of these interactions.

The distributions of PTMs across domains seem to indicate, in some cases, potential steric hindrances and inaccessibility of PTM deposition enzymes to specific RBDs and RNA-binding regions (**Figure 2.2D**). The most evident case is the location of methylation sites relative to RRMs. While these potential sites are barely found within RRMs, methylations seem to be rather clustered downstream of these domains. Mechanistically, these sites could act as switches which induce conformational changes of RRM tandems, subsequently affecting the RBP-RNA interactions. In fact, it has been recently demonstrated that SUMOylation at different lysine residues of the RBP HuR produces differential rearrangements of RRM1 relative to RRM2, modulating its RNA-binding affinity (Lachiondo-Ortega et al. 2024). Indeed, the dysregulation of the methylation-SUMOylation interplay could be a key factor yet to be explored in a context of tumor onset and progression, in some cancer types such as hepatocellular carcinoma (Lachiondo-Ortega et al. 2024). Lastly, our work also evidenced IDRs as densely PTM populated regions (**Figure**

2.2D), which could be explained by a higher degree of surface exposure making them more prone to PTM deposition (Sternburg, Gruijs da Silva, and Dormann 2022).

Moving forward, we addressed the capability of PTM site presence to cluster RBP-RBP pairs. As shown in **figure 2.1,** the PTM site presence was represented as a categorical features binned to 0/1 to indicate absence or presence of sites on the RBP. The unsupervised learning algorithms employed showed well-defined clusters when the PTM features were present in the matrix.

However, we need to consider that combining binary qualitative quantitative features could result in a potentially biased representation. Whereas the quantitative features in the matrix (e.g. the intersection frequencies) are scaled in a 0-1 range, and their values are in most cases intermediate rather than 0 or 1, the binary features values are, by definition, either 0 or 1. It must be taken into account that t-SNE algorithm will interpret those values as extreme cases of quantitative features ignoring its qualitative nature, giving them an unproportional weight on the bidimensional projection. Nevertheless, there is an evident impact of the tested PTMs on RBP-RBP interactive features, reflected on the fold changes observed when comparing the groups of RBP with different PTM status (**Figure 2.7A,B**). Moreover, the comparison between these PTM status revealed significant differences for features as important as the frequencies of intersecting and proximal RNA-binding sites, also suggesting potential shifts on the RNA affinity caused by methylation and acetylation (**Figure 2.7C**).

As it was previously explained in the discussion of chapter 1, RBP-RBP interaction features derived from eCLIP assays manifested several sources of bias, which must also be considered in the present analysis. In particular, the selection of the maximum distance between nearby binding sites as well as the positive correlation observed for intersecting and adjacent binding sites could be potentially influencing our results and the consequent interpretation.

The analysis of RBP interaction networks revealed a significantly higher number of PPIs on RBPs with acetylation sites compared with non-acetylation sites RBPs. In contrast, it was also observed a significant decrease in intersections and nearby binding sites when we compared the group of RBP pairs harboring acetylation sites (in one and both RBPs) against RBP pairs lacking acetylation sites (**Figure 2.7C**). Although these two findings a priori might seem contradictory, a potential explanation could be that acetylation actually promotes PPIs while diminishing the RNA-binding activity of those RBPs. Moreover, this could also imply that acetylation sites may determine a trade-off between PPIs and RNA-protein interactions.

An alternative explanation could be RBP sequestration for this enrichment in PPI and the consequent depletion of RNA-protein interactions, which could be an unknown role of acetylation consistent with the imbalance between these two types of molecular interactions (Berenson et al. 2023; Zhonghua Wang, Bhattacharya, and Ivanov 2015).

Regarding the post-transcriptional control of PTM deposition enzymes, RBPs enriched on this targets seems to have a preference for intronic regions of these enzymes transcripts as can be observed in **figure 2.9B.** In addition the GO enrichment analysis revealed an overrepresentation of terms related to RNA splicing and mRNA processing functions. Taking this into account, our results suggest that RBP occupancy could be the most common mechanism for retention and splicing of such introns, thus shaping the production of transcript variants. Moreover, the greater the involvement in PTM-enzyme targeting, the stronger the enrichment in GO splicing terms, implying an underestimation of this potential RBP-PTM regulatory loop until now.

While the addition and removal of PTMs can lead to a shift in RBP binding preferences ("Spatial Correlation Statistics Enable Transcriptome-Wide Characterization of RNA Structure Binding" 2021; Lachiondo-Ortega et al. 2024), our results regarding whether potential PTM sites could be a cause of differential binding are still preliminary. In this context, experimental validation could shed light on the shifts in RBP activity when this RBP is modified in contrast to a mutated RBP lacking the PTM site. Such precise alterations in protein function are often linked to PTM sites, having significant implications for human health. Gene mutations affecting PTM sites can trigger numerous diseases in humans.

In addition, analogously to how studies focused on individual RBP actions have been progressively replaced by complex regulatory networks, previous works focused on a single PTM type are obsolete as well. The fact that proteins, and particularly RBPs, can harbor multiple potential PTM sites of different types suggests that different PTM mechanisms coexist on the same protein, manifesting a combinatorial regulatory effect (Geffen et al. 2023; Lachiondo-Ortega et al. 2024).

Material & methods

RBP data collection and preprocessing

We used the list of human RBPs defined by (Sebestyén et al. 2016). Four additional RBPs (DGCR8, EXOSC5, LSM11, TROVE2) present in the ENCODE eCLIP dataset and not originally in this list were also included. We used the Uniprot portal ("UniProt: The Universal Protein Knowledgebase in 2023" 2023) to extract, for each RBP, the type and position on the peptide of each PTM, Gene Ontology annotations, and the location of RNA binding domains and regions. We then searched for acetylation, methylation, and phosphorylation in the modified residues annotated for each RBP, and built a one-hot encoded matrix where 0/1 represent absence/presence of a specific PTM.

Analysis of PTM sites position

To compute the relative position of PTM sites in RBPs, each modified residue on the protein was normalized to the 0-1 range, with zero being the N-terminal and one the C-terminal amino acid of the peptide chain. We used the start and end positions of each RBD and RNA binding region to determine the relative position of PTM within those. For PTM sites in proximity, we defined a window of 50 amino acids upstream and downstream the domain or region and we computed their relative position within the window. For the assessment of significant enrichment of sites toward the N or the C-terminal end of the peptide we performed a t-test using the center of the distribution as mean (mu = 0.5) and a confidence interval of 95%.

eCLIP data processing and peak annotation.

eCLIP RNA-binding sites for the batch of 83 ENCODE RBPs from HepG2 cell line were processed as described in material & methods section of chapter 1. Then, the transcript regions mapped for each RNA-binding site were annotated using CTK (Shah et al. 2017). Among the pool of RNA-binding sites, some of them mapped in more than one transcript region due to alternative splicing events, which is reported by CTK as an ambiguous annotation (e.g: CDS||3'UTR). To avoid this ambiguity in further analysis, we exclusively retained the region with the higher priority according to the following order: CDS, 5′UTR, 3′UTR, intron, exon_unclassified.

Prediction of RNA secondary structure elements

RNA secondary structure elements were predicted with EDeN (Navarin and Costa 2017) , providing full transcript sequences as input for transcripts having at least an eCLIP binding site. Binding sites on exons were mapped to the transcriptome using the ensembldb R

package (Rainer, Gatto, and Weichenberger 2019). Secondary structure elements were then mapped to sequences derived from GENCODE (Frankish et al. 2021). For binding sites mapping on introns, the secondary structure of the full intron was predicted and the binding stretch extracted using the GenomicRanges and plyranges R packages (Lawrence et al. 2013; S. Lee, Cook, and Lawrence 2019)). From EDeN's output, we computed the median relative abundance of each structural element over all binding sites of each RBP.

Analysis of RBP subcellular locations

RBP subcellular locations were extracted from the Human Protein Atlas (HPA) (Thul et al. 2017), including both main and additional locations. HPA annotations were then converted to a one-hot encoded matrix where each column was a subcellular location and each row a different RBP with 0/1 representing the absence/presence of that RBP on that location. Enrichment in potential PTM sites at each subcellular location was computed by Fisher's exact test with a threshold of 0.05 on the Benjamini-Hochberg adjusted p-value.

Functional RBP clustering

The functional subgroups of RBPs were identified from a hierarchical clustering performed over the MF and BP terms reported with significant enrichment or depletion in any of the three PTMs on RBPs. After generating RBP clusters with UMAP we inferred the cluster membership of each RBP by neighborhood majority. In the case of ties, we applied the color corresponding to the MF with the highest number of points

Analysis of RBP interaction networks

Experimentally validated protein-protein interactions between RBPs were retrieved from STRING (Szklarczyk et al. 2019). Given the high network interconnectivity, we selected interactions with a minimum confidence score of 400 and we retrieved only one-way edges (A-B), filtering out the not reciprocal ones (A-B B-A), to avoid the excess of edges. We then assigned a color code based on the presence and absence of PTMs on each RBP. STRING data was loaded to R with the igraph package (Csárdi and Nepusz 2006), and previously computed PTM sites annotations were then used to attach the PTM information to the graph as a node attribute defining which RBPs presented sites for acetylation, methylation, or both. Networks were then plotted with Cytoscape (Shannon et al. 2003). We conducted the Shapiro-Wilk test on the number of PPIs per RBP to assess normality and subsequently chose the U-Mann Whitney test for the group comparison.

Additional computation of RBP-RBP features.

Intersection and nearby binding site frequencies were computed as described in the material & methods section of chapter 1. The same features were also computed by subsetting RNA binding sites according to the transcript regions (e.g. 3'UTR) where they mapped and the corresponding transcript biotypes (e.g. protein-coding).

Aside from features describing potential RBP-RBP interactions (those derived from eCLIP), we also included features detailing the intrinsic properties of individual RBPs. This included the presence of PTM sites (i.e. acetylation, methylation and phosphorylation), the protein length, the presence of nuclear localization motifs, and the fraction of IDRs and structural elements (α-helices, β-strands, coils and turns) predicted by AlphaFold (Varadi et al. 2022; Jumper et al. 2021).

XGBoost hyperparameter tuning and model training

As other tree-based models, XGBoost is considerably prone to overfitting when a certain number of boost rounds is exceeded. To avoid overfitting, we performed k-fold cross validation computing different metrics as the number of boost rounds increased. Then we selected through the elbow method the number of rounds that reasonably reduced the error metrics without risking to overfit the model, which was approximately 10 in all cases.

The objective function for the regressor mode was the "reg:squarederror". The evaluation metrics used as reference to select the number of boosting rounds were the residual mean squared error (RMSE) and the mean squared error (MAE). The objective function selected for the multiclass classifier mode was "multi:softmax", with three different classes (i.e. none, one and both) .The evaluation metrics considered in this case were the multiclass error, the multiclass negative log-Íikelihood and the area under the ROC curve (AUC).

XGBoost performs well with the default parameters. However, by tuning the hyperparameters, this performance can be increased up to approximately 5%. To do this, we carried out a grid search with multiple values of each hyperparameter and we scored their performance, eventually finding the best combinations for the classifier and regressor modes (**Supplementary table 1.4**) (T. Chen and Guestrin 2016).

As explained in the previous method section, the set of features were computed first by considering all RNA binding sites without making distinctions and second by subsetting the transcript regions and biotypes where these were mapping. However, these last sets of features were excluded from the XGBoost input data to avoid any possible confounding effect and to reduce complexity in both the regressor and classifier.

RBPs target enrichment analysis

We extracted the lists of genes coding for protein acetylation and methylation enzymes from the HGNC (Seal et al. 2023). We used the IDR set of peaks from the ENCODE eCLIP datasets (Luo et al. 2020) to compute the enrichment of RBP binding sites in these genes. Enrichment was computed by Fisher's exact test with a threshold of 0.05 on the Benjamini-Hochberg adjusted p-value.

Chapter 3: Differential expression of RNA modifier proteins in breast cancer.

Background

BC is the most diagnosed cancer type in women and represents one of the most frequent cancer-related deaths for females (Kumari, Groza, and Aguilo 2021). This cancer type is a heterogeneous disease composed of multiple molecular subtypes which significantly vary in biological features and clinical implications. Breast tumors differing on histopathological and biological characteristics also exhibit distinct behavior, which in turn leads to different treatment responses (Dai et al. 2015). Taking this into account, making accurate classifications into subtypes is fundamental for therapeutic decision making.

Breast cancer subtypes are typically determined through gene expression profiling and immunohistochemistry (IHC). Tissue microarray has been classically performed for subtype determination based on gene signatures. However, many discrepancies on gene signatures have arisen among different studies. Classification methods are mostly based on the expression of specific hormone receptors, i.e. the estrogen receptor (ER) and progesterone receptor (PR), and the human epidermal growth factor receptor 2 (HER2). These IHC markers along with the tumor size, grade and other pathological features are classically used for patient prognosis (Staaf et al. 2022; Dai et al. 2015).

Despite some inconsistencies on gene expression signatures, up to date, five major intrinsic breast cancer subtypes with different clinical outcomes have been widely recognized based on the current PAM50 classification established by (Parker et al. 2009): Normal-like, HER2 overexpression, Luminal A, Luminal B and Basal (also referred to as triple negative breast cancer, TNBC). Each of these intrinsic subtypes is associated with a distinct molecular and clinical profile with the exception of the Normal-like subtype which remains less clearly defined and accounts for approximately 8% of breast cancer cases (Dai et al. 2015; Orrantia-Borunda et al. 2022).

Among all breast cancer subtypes, TNBC is the most heterogeneous and is associated with the worst prognosis. The basal-like tumors commonly referred to as TNBC are characterized by the absence of the three conventional markers, ER, PR and HER2 thereby rendering hormonal therapies ineffective (Yersal and Barutca 2014; Kumari, Groza, and Aguilo 2021). Recent research has further classified TNBC into four molecular subgroups: Basal-like 1, basal-like 2, mesenchymal subtype and luminal androgen receptor (LAR) (Lehmann et al. 2011), each exhibiting distinct biological and clinical features.

The next subtype from worse to better prognosis is the HER2 subtype, characterized by an overexpression of HER2, and the HER amplicon genes (e.g. GRB7 and PGAP3). However, identification techniques such as immunostaining or fluorescence in situ do not match all clinical HER2+ tumors because of the invariability at transcriptional level (Kumari, Groza, and Aguilo 2021; Orrantia-Borunda et al. 2022). Despite showing an overall less aggressiveness than TNBC, the HER2 subtype has a poor prognosis as well due to a higher risk of relapse after eradication of tumor cells, and they are more likely to reach grade 3 (Dai et al. 2015; Orrantia-Borunda et al. 2022).

Lastly, the luminal-like tumors resemble those of the luminal epithelial cells from the breast. These tumors are characterized at molecular level by the expression of hormone receptors (ER+ PR+). Luminal-like breast cancers are subsequently subcategorized into luminal A and B (Dai et al. 2015; Orrantia-Borunda et al. 2022). Whereas Luminal A expression profiles are mainly HER2- , luminal B can be HER2+/- . In general, luminal B tends to be more aggressive and present a higher level of the cell proliferation marker Ki-67 (Yersal and Barutca 2014; Kumari, Groza, and Aguilo 2021).

Multigene expression assays have been classically used as guidance to determine the prognosis of breast cancer patients. However these tests are limited to a number of clinical

useful outputs per analysis due to their targeted design. In this context, transcriptome-wide RNA-sequencing represents a better alternative to conventional subtyping systems based on prognostic multigene classifiers (Staaf et al. 2022).

The assignment of BC samples to their respective molecular subtypes has typically relied on the nearest-centroid (NC) algorithm. The NC assigns a class label to each new sample by measuring the distance in a relative gene expression space to class centroids, thus selecting the nearest one. One major disadvantage of NC is the dependency on previous samples for the normalization of new samples added to the dataset, making it unsuitable for cluster initialization and small datasets. In this context, the work of (Staaf et al. 2022) presents a novel classification method referred as single sample predictor (SSP) which relies on the concept of gene-pair rules. This new classification approach consists in evaluating intrinsic relations between genes within a single sample for the class labeling task (e.g. gene A expression > gene B expression), thus omitting the normalization required for inter-sample comparisons specific to the NC classifier.

Taking this into account, in this study we conducted a differential expression analysis of 49 genes modulating the deposition and removal of RNA modifications, particularly associated with mRNA (or mRNA modifier proteins, mRMPs), across all pairwise comparisons of breast cancer subtypes classified according to the SSP framework within a specific cohort from the Swedish Cancerome Analysis Network - Breast (SCAN-B, **Figure 3.1A**). SCAN-B is a research initiative focused on improving the clinical management of breast cancer through comprehensive genomic analysis (Saal et al. 2015).

SCAN-B comprises information from 7743 patients, each annotated with one of the following breast cancer subtypes (SSP classification): Basal, Her2, Luminal A, Luminal B and Normal-like. As shown in **figure 3.1B**, there is a discrepancy of 15% in the subtype labelling between the SSP and NCN classificators. In addition, a smaller cohort of 66 samples from normal breast tissue were considered as negative control for the analysis. The SCAN-B dataset encompasses clinical features such as risk of recurrence (ROR), relapse and overall survival (OS) alongside molecular markers such as Ki67 levels which reflect cell proliferation activity (**Figure 3.1C,D**),

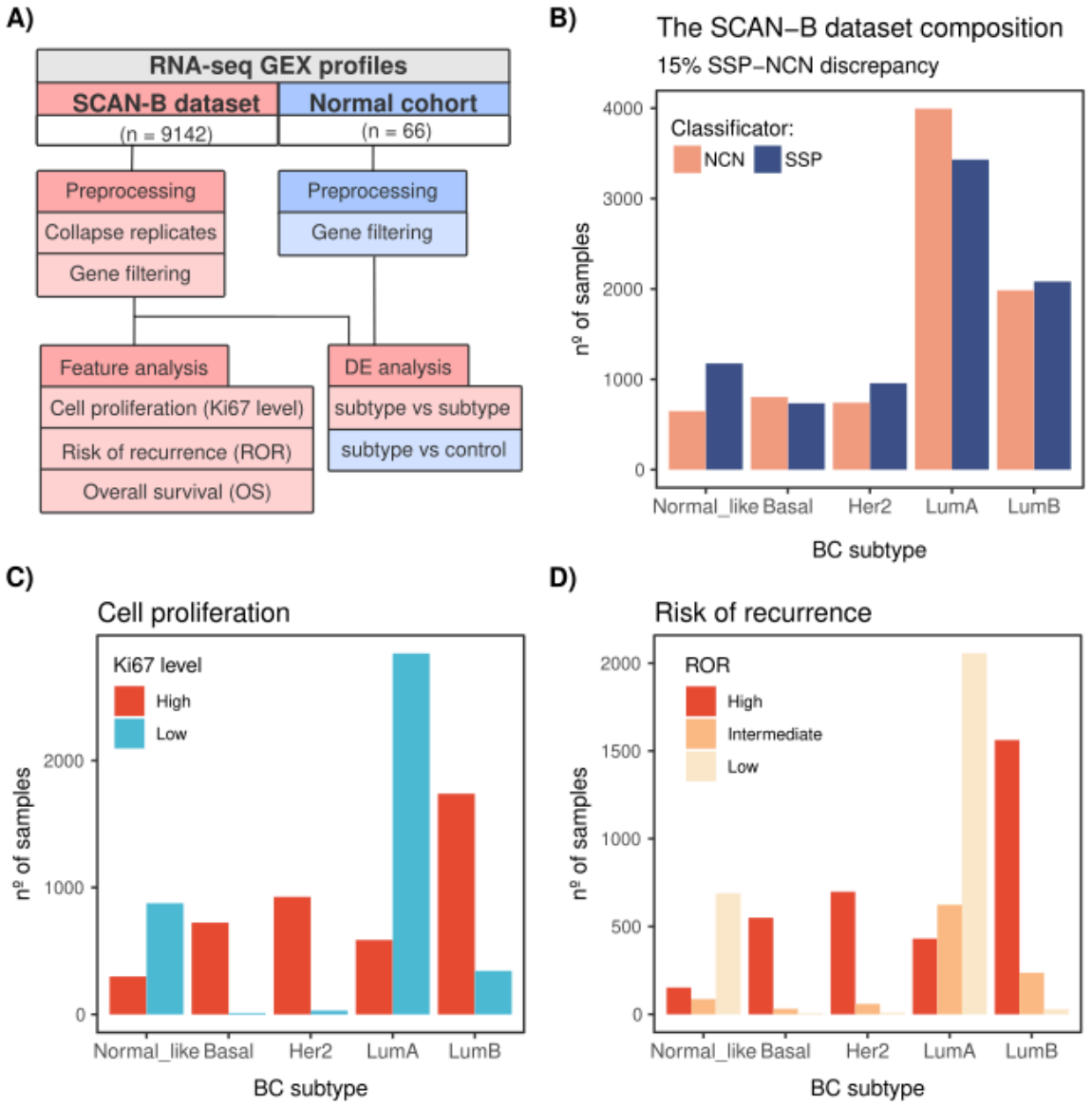


Figure 3.1: Graphical abstract and overview of the SCAN-B cohort features. *(**A**) schematic workflow describing the preprocessing steps and downstream analyses of the SCAN-B data (**B**) Barplots showing the discrepancy in the nº of SCAN-B samples assigned to each BC subtype by NCN and SSP classification methods (orange and blue respectively) (**C**) Cell proliferation rate estimated by Ki67 levels and binned to High/Low for the SCAN-B samples across the different BC subtypes assigned by SSP. (**D**) Risk of recurrence (ROR) expressed as "High", "Intermediate", and "Low" for SCAN-B samples across the different BC subtypes.*

We used this dataset to investigate the differential gene expression of 49 mRMPs. The specific genes are listed in the **supplementary table 2.1**

Results

Differential expression analysis revealed mRNA modifiers upregulated in TNBC.

For the present study we selected a set of 49 mRNA modifier proteins (mRMPs) associated to one the following RNA modifications: m^6A, m^5C, Ψ, 2'-O-Me, C-to-U and A-to-I (**Supplementary table 2.1**). Leveraging the SCAN-B RNA-Seq data, we performed differential expression analysis for all pairwise comparisons of BC subtypes and against the control. To assess significantly down and up-regulated genes across the comparison, we set the cutoff of log2(p-val) > 1.301 and |log2(FC)| > 1. Interestingly, our results revealed a set of 13 mRMPs significantly upregulated in the TNBC subtype against the control, namely IGF2BP1, IGF2BP2, IGF2BP3, NSUN5, ALYREF, YBX1, YBX2, PUS1, MARS, RPUSD2, ADAR, APOBEC3A and APOBEC3G, and the down-regulation of METTL14 (**Figure 3.2A,B**). Overall, the combined overexpression of this set of mRMPs is significantly associated with a worse prognosis of BC patients (p-value < 0.0001, OR = 3.919; **figure 3.2C**). All p-values and fold changes for the rest of mRMPs across the different comparisons are provided in **supplementary table 2.2.**

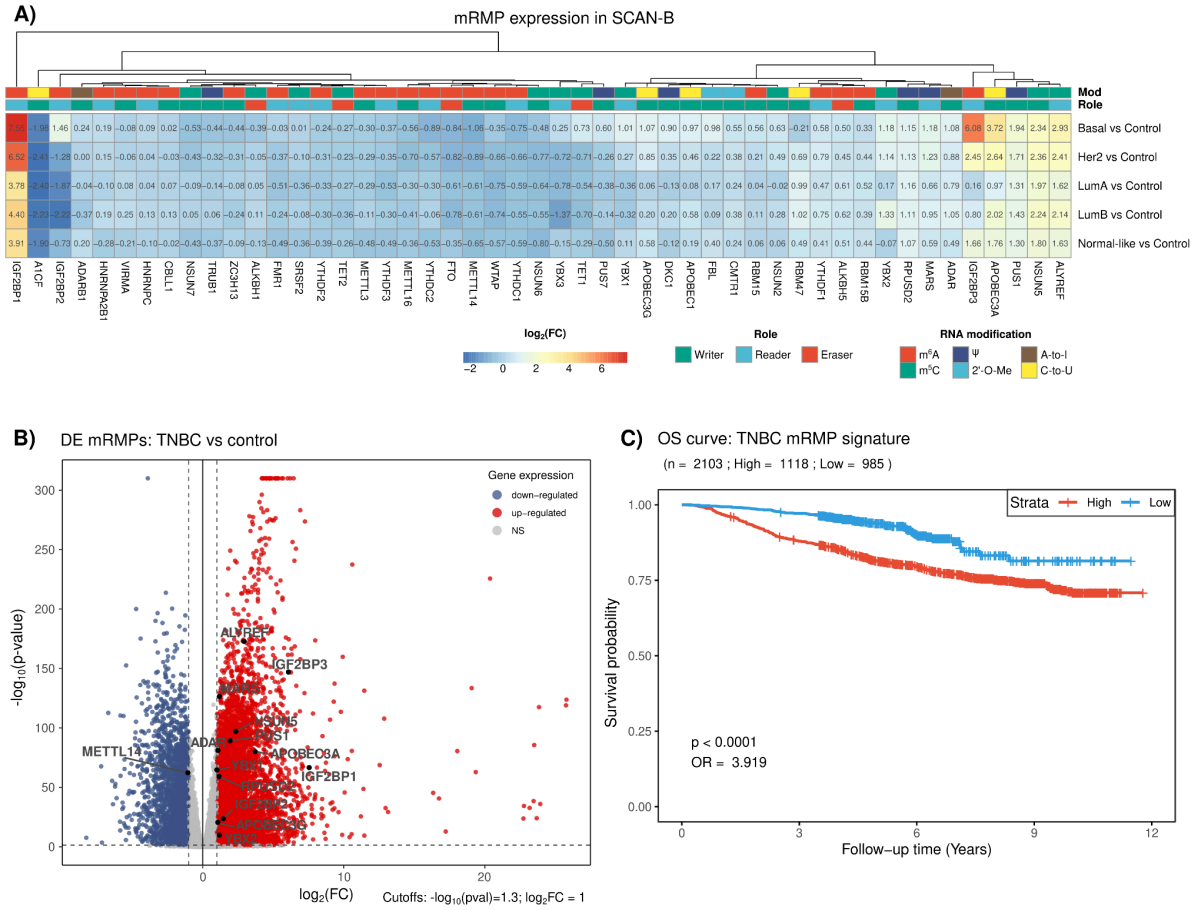


Figure 3.2: Differential expression of mRMPs in TNBC: *(**A**) Heatmap displaying the log_2(FC) values for the differential expression analysis of RNA modifiers across BC subtypes against the cohort of the normal breast samples (control group). Rows represent the different subtype vs control comparisons and columns represent RMPs. The cell value is the log_2(FC) of gene expression between two conditions. RMPs highlighted in red were found up-regulated in the basal type. (**B**) Volcano plot showing differentially expressed genes for the basal subtype vs control comparison. Cutoffs for -log_{10}(p-value), and log_2(FC) were set to 1.301 and 1 respectively. Down-regulated genes are represented in blue and upregulated in red. Only mRMPs over the threshold of significance were labelled (**C**) Overall survival analysis of SCAN-B patients based on TNBC-upregulated mRMPs. The combined effect of the 13 mRMPs was computed by averaging their RNA expression levels. Samples expressing high RNA levels in the majority of genes were included in the high expression group (red line) whereas samples with the majority lowly expressed constituted the low expression group (blue line).*

IGF2BP1, IGF2BP2 and IGF2BP3 are upregulated within TNBC subtype and associated with an overall worse prognosis

The DE analysis revealed the overexpression of the m^6A readers IGF2BP1, 2 and 3 in TNBC vs control (log_2(FC): 7.55 1.46 and 6.08; adjusted p-values:< 0.0001, **figure 3.2A-B**). Given the observed upregulation in TNBC patients, we conducted further analysis of IGF2BP1/2/3 to assess their impact on molecular features and prognosis.

To study potential groups and gene expression across samples we performed PCA for dimensionality reduction and visualized the SCAN-B samples in a 2D space (**Figure 3.3.A**). As expected from our previous results, samples with higher IGF2BP1, 2 and 3 expression mapped on the basal subtype region of the PCA projection (**Figure 3.3A**). Indeed the RNA levels for the three m6A readers resulted significantly higher in the basal subtype (**Figure 3.3C**).

Further examination revealed that out of 734 basal samples, 52.6% expressed high levels of IGF2BP1, 81.9% expressed high IGF2BP2 and 88% expressed high IGF2BP3. To account for samples highly expressing genes across SCAN-B samples, the DESeq2-normalized counts were stratified into high, intermediate, and low expression using quantiles 25th and 75th as lower and upper limits respectively (see **Material & methods**). The degree of overlap between highly expressed IGF2BPs across samples is represented as Euler diagrams in **figure 3.3B** for the full SCAN-B cohort (**left column**) and the TNBC subset (**right column).**

Building on their association with the TNBC subtype, we investigated the impact of IGF2BP1/2/3 on various molecular and clinical features to evaluate their potential influence on prognosis. To determine whether the expression of these genes contribute to increased cell proliferation rates, samples were splitted into two groups according to the Ki67 level, categorized as high and low, and a t-test was conducted for mean comparisons between groups. Notably, samples with high Ki67 protein levels exhibited significantly elevated RNA expressions of IGF2BP1, 2 and 3 (p-value < 0.0001 in all cases; FC: 4.9, 2.1 and 8.6 respectively) (**Figure 3.3D**).

Moving forward, we further explored the impact of IGF2BP2/3 on prognosis by analysing clinical features. In particular, we asked ourselves whether a high expression of these factors could be associated with a high risk of recurrence (ROR). SCAN-B samples were thereby

splitted according to high and low ROR and a t-test was performed to compare means of gene expression levels. Both IGF2BP1 and 3 showed significantly higher expression in patients with high ROR compared to the low ROR group (p-values adjusted < 0.0001 in both cases, FC: 13.75 and 1.96 respectively) (**Figure 3.3E**)

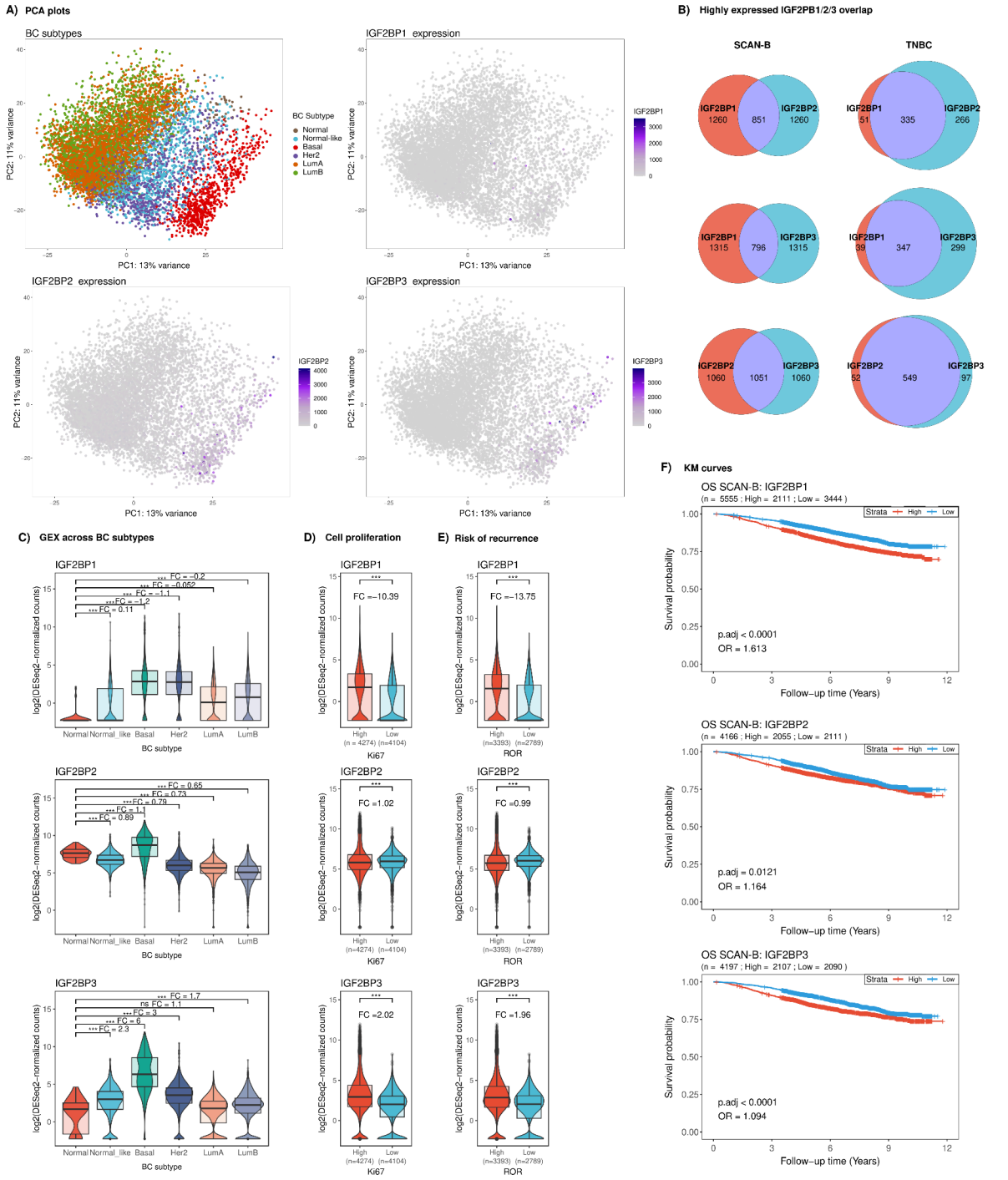


***Figure 3.3: IGF2BP1/2/3 expression is associated with TNBC disease**. (**A**) PCA of SCAN-B samples based on their RNA-seq profiles. Each point on the scatter plot represents an individual sample. The color code on the upper left plot corresponds to the BC subtype assignment performed by SSP classification. The other plots correspond to the gene expression of IGF2BP1, 2 and 3 respectively and the color gradient reflects the RNA expression level across samples (DESeq2-normalized counts) (**B**) Euler diagrams displaying the overlap between high-expressed IGF2BP1/2/3 samples in the SCAN-B dataset (left column) and the subset of TNBC samples (right column). Samples are annotated as "High expression" for a specific gene when the DESeq2-normalized count was higher than quantile 75th (**C**) boxplot set of SCAN-B DESeq2-normalized gene counts for IGF2BP1/2/3 across the different BC subtypes and the cohort of normal breast samples. (**D**) Cell proliferation boxplots. Mean comparison for IGF2BP1/2/3 expression between samples with high and low Ki67 levels. (**E**) Risk of recurrence boxplot. Mean comparison for IGF2BP1/2/3 expression between samples with high and low ROR (**F**) Overall survival analysis for IGF2BP1/2/3. Kaplan-Meier (KM) curves were constructed with stratified data based on quantiles 25th and 75th of gene expression (i.e. Low if < q25 and high if >q75) . High group is represented in red and low group in blue. The p-values correspond to the log-rank test with multiple test corrections employing the BH method.*

Finally, to assess the impact of IGF2BPs on the overall survival (OS) in BC we used their stratified expression and the clinical data for death events and follow up time on SCAN-B to fit the Kaplar-Meier (KM) model. The OS analysis reveals that patients with high expression levels of IGF2BP2 and IGF2BP3 exhibit significantly poorer survival outcomes compared to those with low expression levels (p-values: 0.001 and < 0.0001 ; OR: 1.29 and 1.37 respectively). The survival curves show that the high-expression group lies consistently below the low-expression group, with statistically significant differences, thus suggesting a potential association of high expression of the IGF2BPs to a worse prognosis of BC patients (**Figure 3.3F**). To complement the information obtained from SCAN-B, an additional OS analysis was performed using the METABRIC dataset from TCGA (Rueda et al. 2019; Pereira et al. 2016; Curtis et al. 2012). In this case, the log-rank test yields significant differences between high and low expression groups only for IGF2BP2 (p-value = 0.0012, OR = 1.73 ; **Figure 3.4B**).

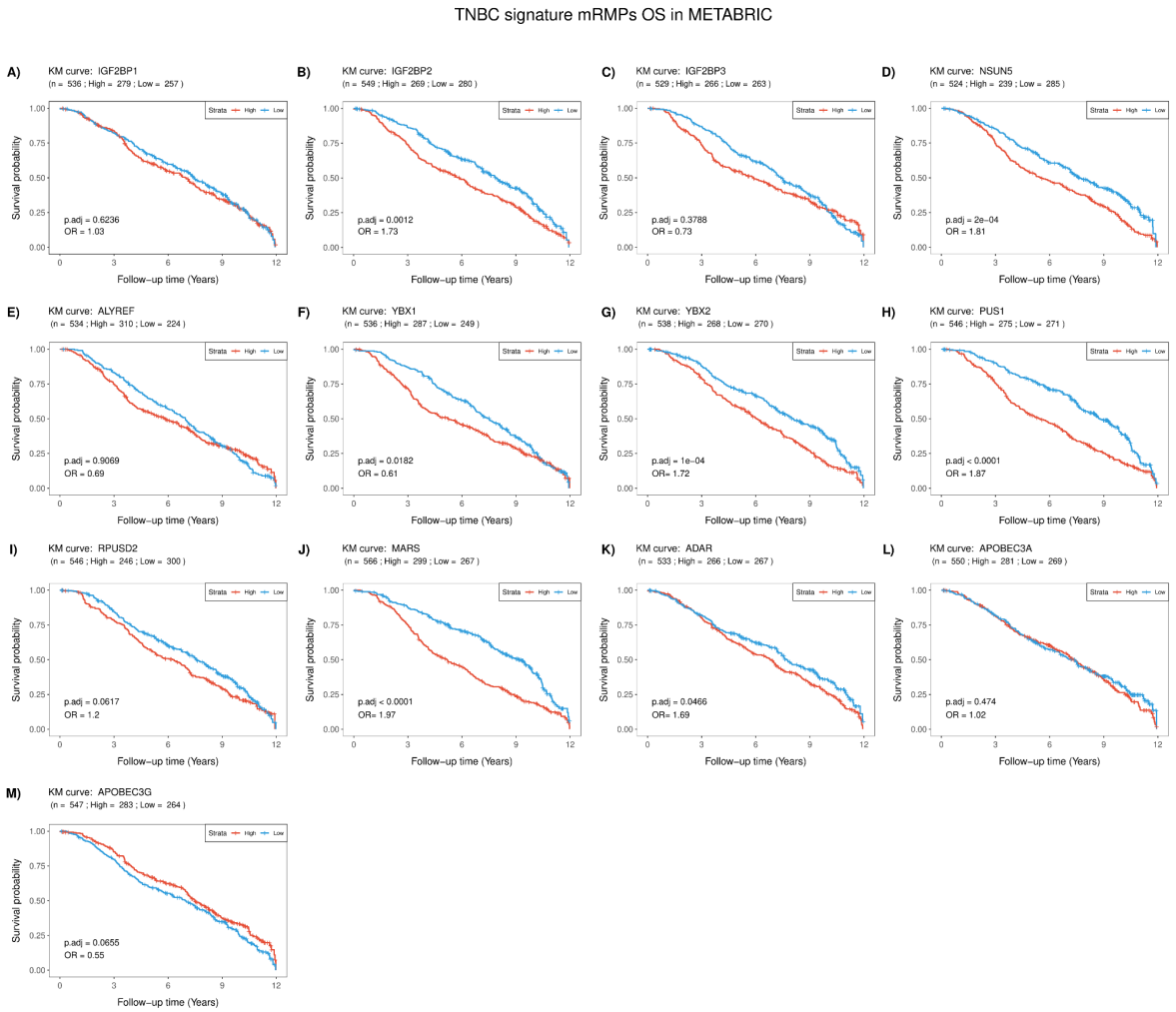


Figure 3.4: Overall survival analysis conducted in the METABRIC dataset for all mRMPs overexpressed in TNBC*. (**A-M**) KM curves were constructed with stratified data based on quantiles 25th and 75th of gene expression (i.e. Low if < q25 and high if >q75) . High group is represented in red and the low group in blue. The p-values correspond to the log-rank test with multiple test corrections employing the BH method.*

The Ψ regulators PUS1, MARS and RPUSD2 contribute to the TNBC-mRMP signature.

We assessed the expression of mRMPs involved in the pseudouridylation of mRNA in the SCANB cohort, and found that the expression of PUS1, RPUSD2 and MARS was upregulated in TNBC compared to normal breast tissues (**Figure 3.5A,C**). The elevated expression of PUS1, RPUSD2, and MARS in BC cases showed a positive correlation with high Ki67 levels (**Figure 3.5D**), suggesting their contributing role in tumor growth. The expression correlation analysis revealed that PUS1, RPUSD2, and MARS expression was

also associated with an increased ROR (**Figure 3.5E**). Notably, higher expression of PUS1 and MARS was significantly linked to reduced overall patient survival, while RPUSD2 expression showed no significant correlation with patient outcomes (**Figure 3.5F**). Consistent with SCAN-B cohort findings, METABRIC analysis showed that high expression of PUS1 and MARS was linked to reduced overall patient survival (**Figure 3.4H,J**). Although elevated RPUSD2 expression suggested a trend toward reduced survival, the results were not statistically significant (**Figure 3.4I**). Altogether, these findings suggest a strong association of PUS1, MARS, and RPUSD2 with an oncogenic phenotype.

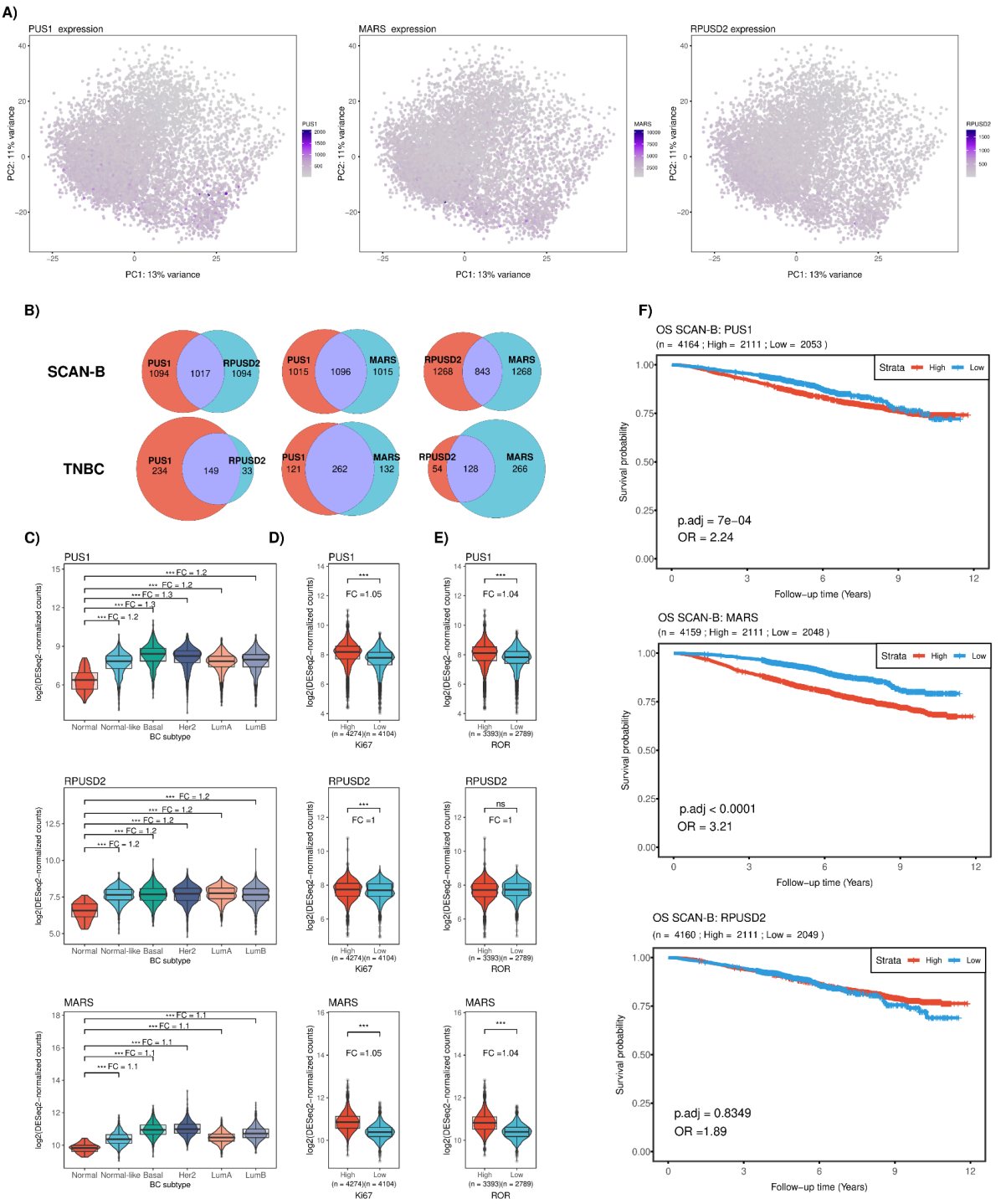


Figure 3.5: Evaluation of RNA expression levels for pseudouridine interactors PUS1, RPUSD2 and MARS. ***(A)*** *PCA of SCAN-B samples based on their RNA-seq profiles. Each point represents an individual sample. The plots correspond to the gene expression of PUS1, RPUSD2 and MARS.The color gradient reflects the RNA expression level across samples (DESeq2-normalized counts).* ***(B)*** *Euler diagrams representing the overlap between samples expressing high RNA levels in SCAN-B cohort (left column) and the TNBC subset (right column).* ***(C)*** *boxplot set of SCAN-B DESeq2-normalized gene counts for across the different BC subtypes and the cohort of normal breast samples.* ***(D)*** *Cell proliferation*

*boxplots. Mean comparison between SCAN-B samples with high and low Ki67 levels. (**E**) ROR boxplot. Mean comparison between samples with high and low ROR (**F**) Overall survival analysis for IGF2BP1/2/3. KM curves were constructed with stratified data based on quantiles 25th and 75th of gene expression (i.e. Low if < q25 and high if >q75) . High group is represented in red and the low group in blue. The p-values correspond to the log-rank test with multiple test corrections employing the BH method.*

The m^5C regulators NSUN5, ALYREF, YBX1 and YBX2 were found overexpressed in TNBC.

Comparison of the expression patterns of the m^5C machinery across breast cancer subtypes revealed that the m^5C methyltransferase NSUN5 and the readers ALYREF, YBX1, and YBX2 were overexpressed in TNBC relative to normal tissue (**Figure 3.6A,C**). Overlaps between samples expressing high RNA levels of such m^5C factors are shown in **figure 3.6B**.

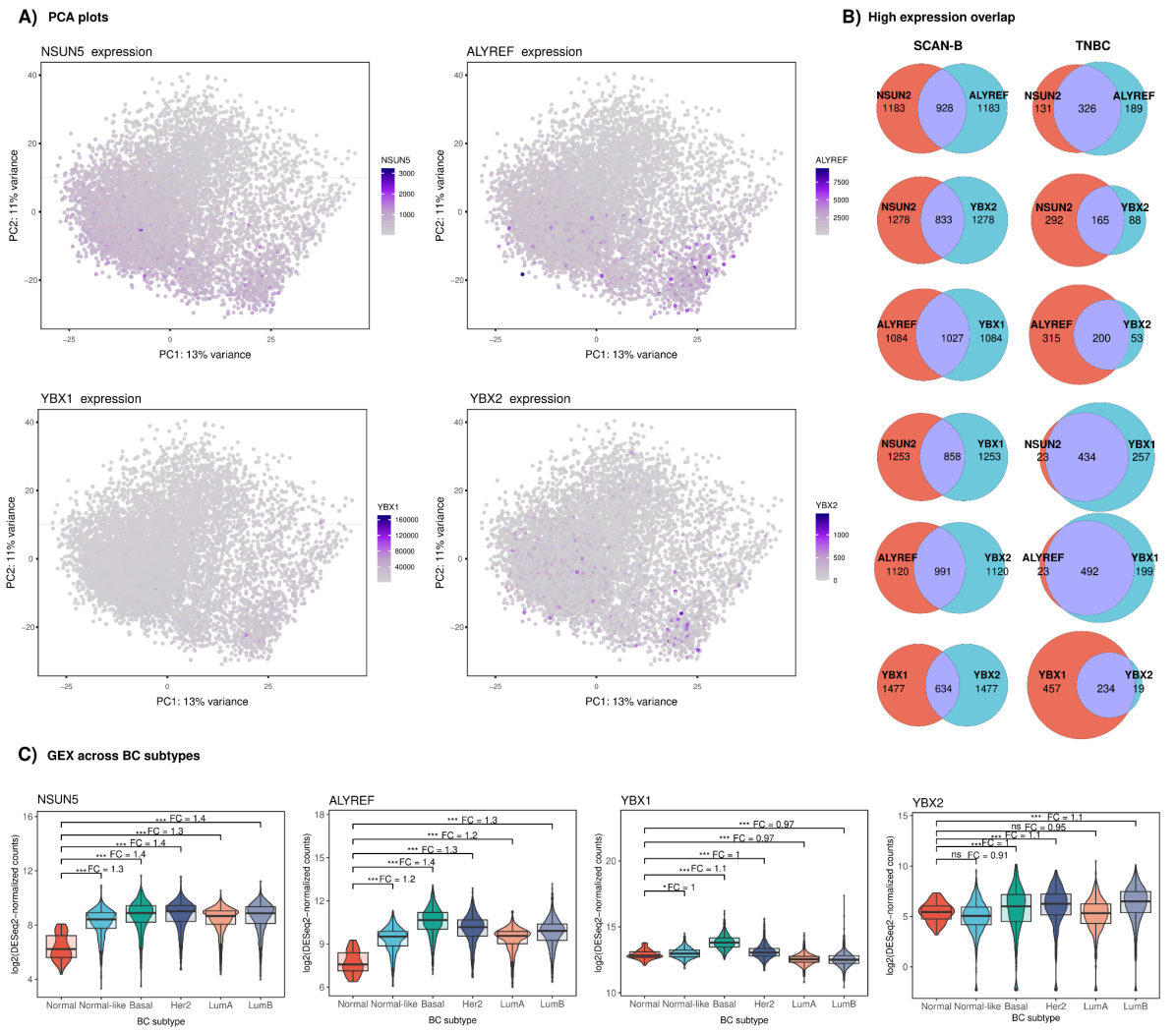


Figure 3.6: Analysis of TNBC-overexpressed m^{5}C factors in SCAN-B cohort. *(**A**) PCA of SCAN-B samples based on their RNA-seq profiles. Each point represents an individual sample. The plots correspond to the gene expression of NSUN5, ALYREF, YBX1 and YBX2. (**B**) Euler diagrams representing the overlap between samples expressing high RNA levels of the m^{5}C factors NSUN5, ALYREF, YBX1 and YBX2 first for the full SCAN-B cohort (left column) and second for the TNBC subset (right column) (**C**) Boxplots showing the comparison between the different BC subtypes for the RNA levels of m^{5}C factors .*

Worse overall survival in the SCAN-B and METABRIC datasets was associated with elevated NSUN5, YBX1, and YBX2 expression, while no clear relationship was found between ALYREF levels and OS (**Figure 3.7A**).TNBC cases characterized by high NSUN5, ALYREF, YBX1, and YBX2 expression exhibited enhanced proliferation rates (**Figure 3.7B, Ki67 plots**) and were associated with an increased ROR (**3.7.B, ROR plots**). Altogether, our analysis highlights the potential oncogenic role of NSUN5, ALYREF, YBX1, and YBX2 in breast tumorigenesis.

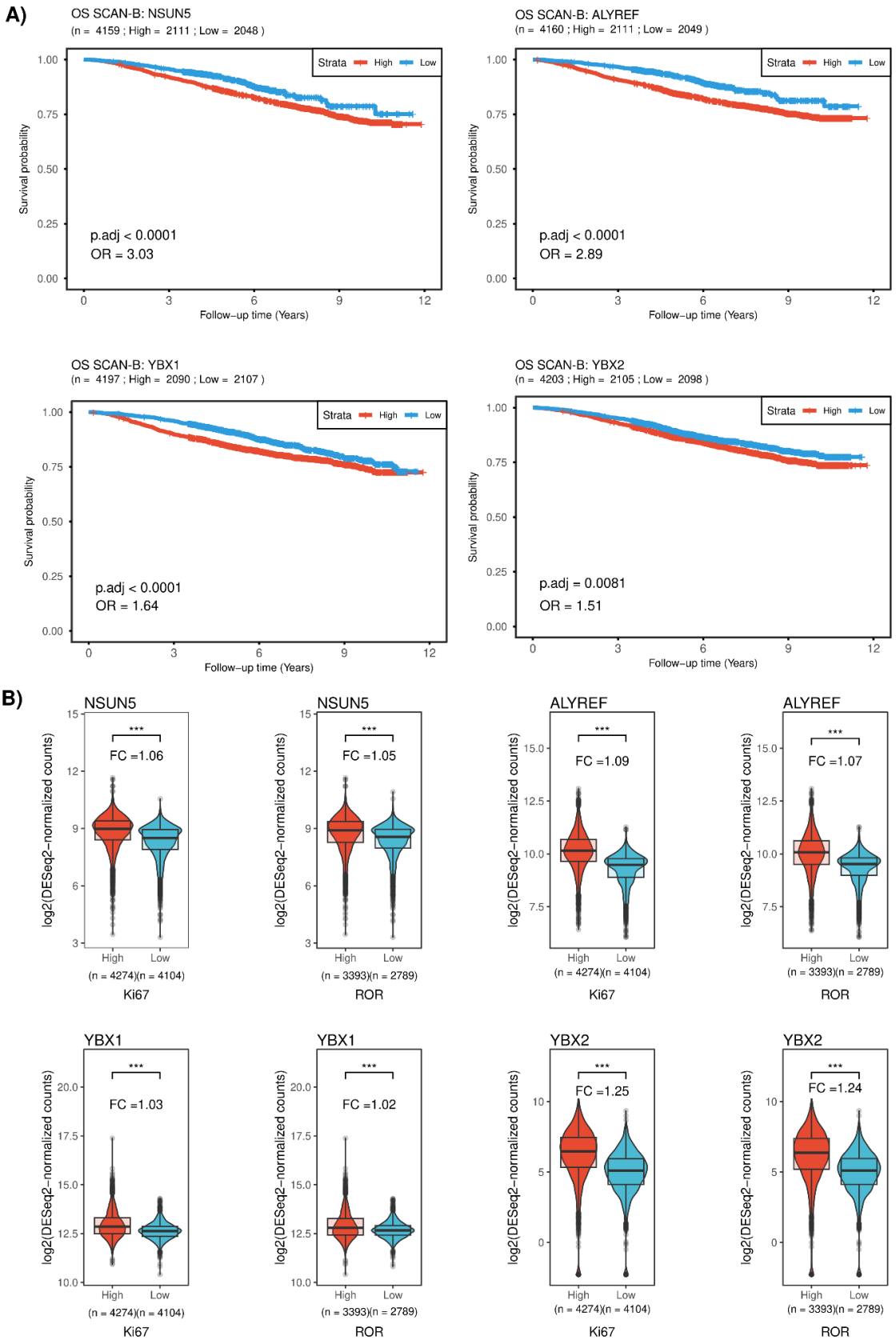


Figure 3.7: Evaluation of clinical features for m^{5}C factors. *(**A**) Overall survival analysis*

*conducted for m^5C factors in the SCAN-B cohort. KM curves were constructed with stratified data based on quantiles 25th and 75th of gene expression (i.e. Low if < q25 and high if >q75) . High group is represented in red and the low group in blue. The p-values correspond to the log-rank test with multiple test corrections employing the BH method. (**B**) Cell proliferation and ROR boxplots. Mean comparison for m^5C mRMPs between samples with high and low Ki67 levels or ROR.*

The DE analysis revealed TNBC-upregulated RNA editing enzymes: ADAR, APOBEC3A and APOPEC3G.

The analysis of the RNA editing machinery revealed an upregulation of A3A, A3G, and ADAR in TNBC compared to normal tissues (**Figures 3.2A,B and 3.8A,C**). Notably, samples with high proliferation rates exhibited significantly elevated expression of these RNA editing enzymes (**Figure 3.8D**). Likewise, patients at high ROR showed increased expression of A3A, A3G and ADAR (**Figure 3.8E**). However, the overall survival analysis revealed a significant association between high expression and a poorer prognosis only for A3A in the SCAN-B cohort (p-adj < 0.0001, OR = 1.723 ; **figure 3.8F**). In contrast, neither A3A and A3G manifested any significant effect in survival probability in the METABRIC dataset, in this case, only ADAR showed a negative impact on prognosis (p-value = 0.0466 OR = 1.69, **Figure 3.4K**). In summary, our findings on SCAN-B are consistent with previous studies that described the strong association between dysregulated expression of RNA-editing enzymes and breast cancer onset and progression.

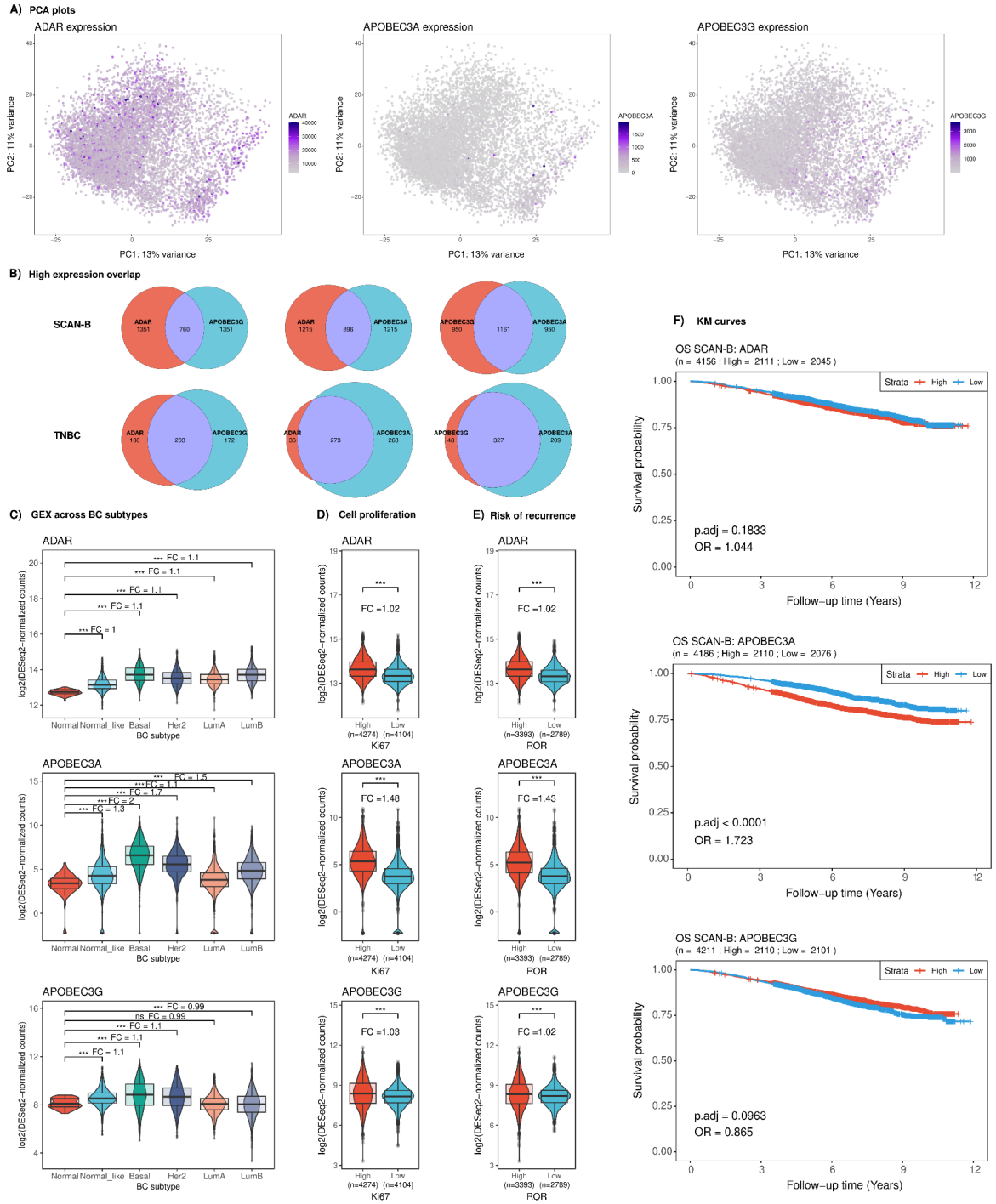


Figure 3.8: RNA expression of ADAR, APOBEC3A, and APOBEC3G in SCAN-B samples. *(**A**) PCA of SCAN-B samples based on their RNA-seq profiles. Each point on the scatter plot represents an individual sample. The plots correspond to the gene expression of ADAR, A3A and A3G respectively and the color gradient reflects the DESeq-normalized count RNA expression level across samples (DESeq2-normalized counts) (**B**) Euler diagrams displaying the overlap between high-expressed ADAR,A3A andA3G samples in the SCAN-B dataset (upper row) and the subset of TNBC samples (lower row). Samples are annotated as "High expression" for a specific gene when the DESeq2-normalized count was*

*higher than quantile 75th (**C**) boxplot set of SCAN-B DESeq2-normalized counts for ADAR, A3A and A3G across the different BC subtypes and the cohort of normal breast samples. (**D**) Cell proliferation boxplots. Mean comparison for ADAR/A3A/A3G RNA expression between samples with high and low Ki67 levels. (**E**) Risk of recurrence boxplot. Mean comparison for ADAR/A3A/A3G expression between samples with high and low ROR (**F**) Overall survival analysis for ADAR/A3A/A3G. Kaplan-Meier curves were constructed with stratified data based on quantiles 25th and 75th of gene expression (i.e. Low if < q25 and high if >q75) . The p-values correspond to the log-rank test with multiple test corrections with the BH method.*

The interaction analysis reveals strong co-occurrent and mutually exclusive mRMP pairs within the TNBC subtype.

The previous sections addressed the effect of different mRMPs on BC features individually. However, given the high complexity and interconnection of the post-transcriptional regulatory network, crosstalk between mRMPs involving different modifications are more than likely. To understand how these factors might interact, further analyses were conducted with the aim of identifying potential interplays occurring among different RNA modifications.

Two different tools were employed to assess co-occurrence (CO) and mutual exclusivity (ME) between mRMP pairs: SELECT and DISCOVER (see **Materials & methods**). The resulting interactive matrices revealed significant CO for m^5C-associated mRMPs and Ψ writers. In particular, NSUN5, ALYREF, YBX1 and YBX2 were significantly co-occurrent, not only among them, but also with PUS1 (**Figure 3.9 A,B**). Complementary to this information, down-regulation of those factors was also found to be mutually exclusive for upregulation of the others. The associations previously described were consistent for both tools tested, strengthening our findings. In addition, another significant CO association was reported for ADAR APOBE3A and APOBEC3G (**Figure 3.9 A**). However, in this case, no significant CO was found by the DISCOVER tool (**Figure 3.9 B**) .

Pairwise Spearman correlation scores were computed for the subset of TNBC-overexpressed genes, first across the full set of SCAN-B samples and then within the TNBC subtype revealing an overall positive correlation among these genes (**Figure 3.9C-D**). Overall, the correlation patterns observed within the TNBC subset of samples resembled

those of the full SCAN-B cohort. The positive correlations observed for m^{5}C regulators and PUS1 reinforce the potential interplay indicated before.

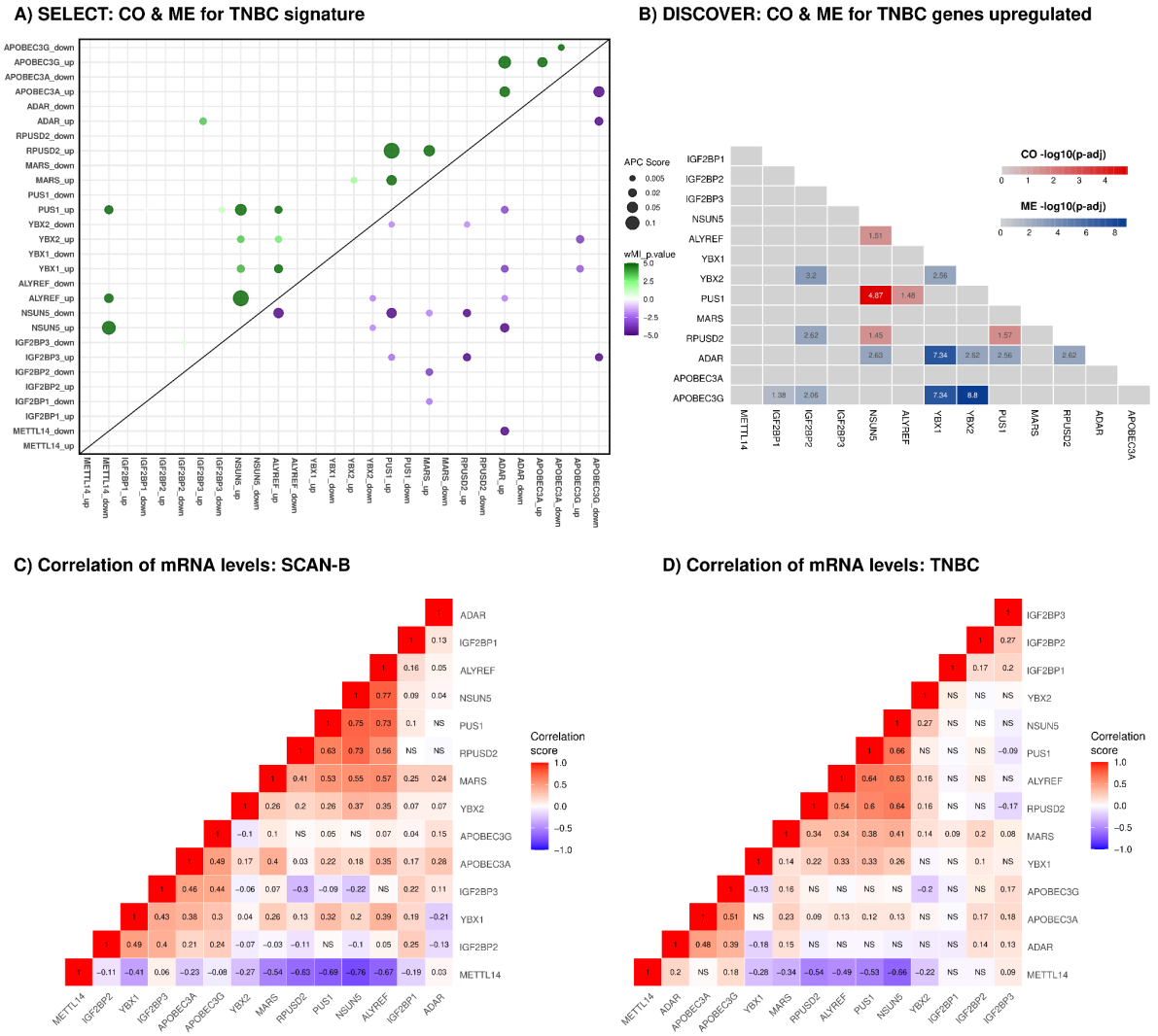


Figure 3.9: Assessment of potential interactions between mRMPs within the TNBC subtype *(**A**) Analysis of co-occurrence and mutual exclusivity between mRMPs employing SELECT algorithm. The color gradient represents the weighted Mutual Information (wMI) p-value which indicates the type of interaction, ranging from ME (violet) to CO (green), whereas the bubble size represents the Average Sum Correction (APC) effect size (**B**) Heatmap of significant CO/ME interactions between upregulated mRMPs exclusively, according to the DISCOVER tool results. Significant CO and ME are represented in red or blue respectively (**C-D**) Spearman correlation scores between mRMPs overexpressed in TNBC. The correlation was computed for the whole SCAN-B cohort (left matrix) and the TNBC subset samples of n = 734 (right matrix). The color gradient ranges represent the degree of correlation and range from -1 (blue) to +1 (red).*

Discussion

IGF2BPs have been reported as enhancers of mRNA stability, playing an antagonistic role to YTHDF2, a well-characterized promoter of mRNA decay (H. Huang et al. 2018; X. Wang et al. 2014). Moreover, an increasing amount of studies highlight the contribution of IGF2BP1, 2 and 3 to fundamental processes in cancer biology, and their overexpression has been associated with poor clinical outcomes in cancer diseases (Mancarella and Scotlandi 2019).

Taking this into account, the overexpression observed for IGF2BP1/2/3 in TNBC suggests a cellular environment where m^6A-mediated mRNA stabilization predominates over degradation. Consequently, m^6A-mRNAs are prone to remain more time being actively translated prior to their degradation which, in turn, could have a potential effect over the imbalance observed in RNA expression levels of other genes. However, IGF2BPs perform m^6A-mediated stabilization of transcripts, which cannot be extrapolated to an extended stabilizer effect throughout the transcriptome. In this context, IGF2BP1/2/3 upregulation is likely associated with specific regulation of a subset of m^6A-marked mRNAs. Moreover, the influence of additional non-m^6A-related factors involved in mRNA stability was not considered for this analysis. Nevertheless, our results, although inconclusive, might indicate a compromised m^6A-mRNA turnover ratio in TNBC cancer cells.

In this context, our findings are consistent with previous studies describing IGF2BPs as stabilizers of oncogenic mRNAs such as c-MYC, KRAS and CTNNB1 (β-catenin) (Elcheva et al. 2023; Die Li et al. 2024). This association to oncogenes could also explain why higher expression of IGF2BPs is found in the group with increased cell proliferation rates.

Elevated pseudouridine levels have been correlated with advanced cancer stages and poor therapeutic response, with increased levels detected in biological samples from various malignancies (Stockert et al. 2021; Zeleznik et al. 2020; Jin et al. 2020). Dysregulated expression of PUS enzymes has been linked to tumorigenesis and cancer progression. Depletion of PUS1 and MARS has been shown to suppress cancerous phenotypes in hepatocellular carcinoma (HCC), renal cell carcinoma (RCC), and BC (Jin et al. 2020; Y.-X. Hu et al. 2024; L. Li et al. 2023). However, the role of RPUSD2 in cancer progression remains unexplored. Consistent with findings in breast and colorectal cancer, we observed increased expression of PUS1, MARS, and RPUSD2 in the SCAN-B cohort (Z. Fang et al. 2022; Kim et al. 2017). These results suggest a potential role for PUS1, MARS, and

RPUSD2-mediated pseudouridine modifications in breast cancer pathogenesis, warranting further investigation.

Tumor cells may promote oncogenic expression through m^5C, as suggested by the upregulation of its readers -YBX1, YBX2, and ALYREF- which enhance mRNA transport, stability, and translation. Abnormal expression of YBX1 and ALYREF, and the m5C writer NSUN5 has been linked to poor survival, though the role of m^5C in these cases remains underexplored. Additionally, since NSUN5 can modify both mRNA and rRNA, linking specific phenotypic effects to individual RNA modifications is challenging, further complicating functional interpretations.

A-to-I editing can alter splicing regulatory elements, leading to aberrant splicing in numerous cancers, while also affecting mRNA stability, localization, translation, and interactions with RBPs, ultimately affecting cancer progression (Fumagalli et al. 2015; Gannon et al. 2018; Gassner et al. 2021). Consistently with previous studies, ADAR was overexpressed in TNBC and high ADAR levels were associated with worse clinical outcomes (Hwang et al. 2019). Additionally, APOBEC enzymes, particularly A3A, have been implicated in cancer progression by inducing mutations that drive genome instability and tumor heterogeneity (Van Norden et al. 2024; Petljak et al. 2022). Abnormal A3A expression has been reported in BC, where it correlates with higher levels of APOBEC-signature mutations (Jalili et al. 2020).

Overall, the results from the interaction analysis suggest the existence of gene blocks which might share common pathways of upregulation under certain oncogenic contexts, for example the one constituted by NSUN5, ALYREF, YBX1, YBX2 and PUS1 or the block of ADAR APOBEC3A and APOBEC3G.

Although positive correlations between mRMPs of the same RNA modification were abundant, we could also observe interesting strong correlated pairs involving different modifications. For example for PUS1 and NSUN5, writers of Ψ and m^5C respectively, suggesting a potential interplay between both modifications (**Figure 3.9C-D**). However, further analysis needs to be conducted to conclude a potential association between the m^5C and Ψ mRMPs.

Nevertheless, it is important to acknowledge the limitations of this study. Gene expression was assessed exclusively at the transcriptional level in SCAN-B samples, which, while indicative of corresponding protein expression, does not provide definitive evidence. Therefore, protein expression levels remain to be validated in this dataset to strengthen our

findings. Additionally, while our analysis reveals significant associations, correlation does not imply causation. Further validation is required to determine whether these relationships reflect direct regulatory interactions or are influenced by confounding factors.

In conclusion, the present work identifies novel mRMPs potentially involved in BC onset and progression, offering new insights into the molecular mechanisms of the disease. By highlighting these underexplored mRMPs, our findings pave the way for future research that could enhance BC diagnostics and treatment, underscoring the critical role of RNA modifications in breast tumorigenesis.

Material & methods

Selection of RNA modifiers with confirmed mRNA-binding function.

In the present study, we considered a total of 49 RNA modifiers with confirmed binding to mRNA species among their pool of RNA targets. The m^6A interactors included all known components of the m^6A writer complex, namely METTL3, METTL14, RBM15, RBM15B, VIRMA, WTAP, ZC3H13, and CBLL1 and METTL16 writer, which has independent methyltransferase activity and high target specificity, mostly catalyzing the methylation of U3 spliceosome RNAs. The m^6A erasers (FTO and ALKBH5) and m^6A readers (YTHDC1, YTHDC2, YTHDF1, YTHDF2, YTHDF3, IGF2BP1, IGF2BP2, IGF2BP3, HNRNPA2B1, HNRNPC, and FMR1) were also included.

Following the same criteria of mRNA targeting, we extended our analysis to RBPs associated with other relevant RNA modifications and RNA editing enzymes in breast cancer: m5C, Ψ, 2'O-Me, and C-to-U and A-to-I. For m5C methylation, we included the known writers NSUN2, NSUN5, NSUN6 and NSUN7, along with the m5C reader ALYREF and additional mRNA-binding proteins YBX1, YBX2, YBX3 and SRSF2, associated with m5C-mediated mRNA processing and stability. Additionally, TET1, TET2, and ALKBH1 were incorporated as m5C erasers or regulators involved in m5C demethylation.

For Ψ, we considered the enzymes PUS1, PUS7 and TRUB1, along with DKC1 and RPUSD2, known to catalyze pseudouridylation in mRNA. MARS (metRS) was also included as Ψ reader due to its association with Ψ-modified transcripts. In the

context of 2’-O-Me, we selected FBL, a core component of the small nucleolar ribonucleoprotein (snoRNP) complex responsible for 2’-O-Me deposition. And CMTR1, a cap-specific 2'-O-methyltransferase involved in mRNA cap structure modifications. Regarding RNA editing. We considered ADAR and ADARB1 (or ADAR2), which catalyze the A-to-I editing, and the C-to-U editing enzymes APOBEC1, APOBEC3A and APOBEC3G, as well as the cofactors RBM47 and A1CF.

This comprehensive set of mRMPs enabled a systematic analysis of the interplay between mRNA modifications, editing and RBP-mediated regulation in the context of BC. The full list of RBPs can be found at **Supplementary table 2.1**.

Differential expression analysis for pairwise comparisons of BC subtypes and against control in SCAN-B dataset

The differential expression analysis across pairwise subtype comparisons was carried out with the DESeq2 R package (Love, Huber, and Anders 2014) . Technical replicates of SCANB samples were collapsed according to the sample ~ merged association found in the metadata provided as supplementary data table (Vallon-Christersson 2023). Low expressed genes with less than 10 counts in total were excluded from the DE analysis. To make the gene counts comparable across samples, we applied normalization by size factor to adjust for the differences in sequencing depth and library size between samples. All breast cancer subtypes were contrasted against the negative control (healthy breast tissue) for the DE analysis. The control consisted of 66 GEX profiles derived from normal breast tissue samples obtained from mammoplasty surgery from healthy women (Staaf et al. 2022). The specific log2FC and p-values for each mRMP are provided in the **supplementary table 2.2**.

Dimensionality reduction and sample visualization.

RNA-seq data for SCAN-B breast cancer samples were combined with the cohort of normal breast tissue samples (control) for joint display. Principal Component Analysis (PCA) was performed for dimensionality reduction for 2D visualization of SCAN-B samples. For this analysis we took into account that highly expressed genes can dominate the PCA (and

UMAP) projections masking the effect of less abundant but biologically relevant genes. To prevent this bias, the gene count matrix was normalized using the variance stabilizing transformation prior to applying PCA and UMAP.

Statistical analysis of RNA-seq data.

Statistical analyses were conducted for mean pairwise comparisons between different BC subtypes and conditions (i.e. Ki67 and ROR; High vs Low). To this end, we conducted multiple two-tailed t-tests for independent samples with a p-value cutoff of 0.05. The p-values were adjusted for multiple testing employing the Benjamini-Hochberg (BH) method. The corrected p-values along with the FC for all statistical tests are shared in the **supplementary table 2.3**. DESeq2-normalized counts were logarithmic transformed for visualization of boxplots.

Survival analysis in the SCAN-B dataset

Evaluation of the prognostic effect of genes in breast cancer was carried out according to the overall survival (OS) time and events annotated in the SCAN-B supplementary table containing the samples metadata (Vallon-Christersson 2023). In parallel, we performed a second OS analysis of the METABRIC dataset for BC available in the cBioPortal (Pereira et al. 2016)

The OS time in days and the death status of the patient binned to 0/1 were employed to construct the Kaplan-Meier (KM) model with the survival R package (v3.5.7), while OS plots were generated with survminer (v0.5.0) (June, n.d.; Ito and Murphy 2013). For each survival curve, samples were stratified according to gene expression. The 25th and 75th quantiles were selected to segregate gene expression into high (GEX > q75), intermediate (> q25 & =< q75), or low (=< q25) samples. The FDR method was employed for multiple test correction of p-values obtained from the log-rank test. The OR computed for each KM curve in SCAN-B cohort and METABRIC can be found in **supplementary table 2.3**.

The combined effect of TNBC-upregulated genes in the overall survival was assessed by averaging their RNA expression levels. Samples expressing high RNA levels in the majority of genes were included in the high expression group whereas samples with the majority of genes lowly expressed constituted the low expression group.

Identification of CO and ME between up- and downregulation of mRMPs in TNBC samples.

To assess the presence of alterations, we computed the FC between the DESeq2-normalized counts of each TNBC sample and the mean expression of normal samples. Only samples with |log2FC()| > 1 were considered as altered for each of the RMPs tested (upregulated, if log2FC > 1, and downregulated, if log2FC < -1), thus building a binary matrix of 1 and 0 representing altered and unaltered samples respectively for each factor.

The analysis of CO and ME was carried out employing two different statistical frameworks: SELECT v1.6 (Mina et al. 2020) and DISCOVER v0.9.4 (Canisius, Martens, and Wessels 2016). Both tools evaluate pairwise alteration patterns by comparing observed data against a randomized null mode to determine statistical significance. The wMI p-values of SELECT were considered negative for ME mRMP pairs and positive for CO. Only interactions under the FDR cutoff and the Average Sum Correction (ASC) effect size threshold were considered significant.

Chapter 4: Additional contributions

Leveraging Oxford Nanopore technology to study the interplay between APOBEC-induced A-to-I editing and m^{6}A modification.

RNA editing is a post-transcriptional mechanism which can modify transcripts through indels and base substitutions. RNA-editing occurs in a wide range of organisms, and is involved in several biological functions. There are two types of RNA-editing events in mammals, adenosine-to-inosine (A-to-I) which is catalyzed by ADAR family of enzymes, and cytosine-to-uridine (C-to-U), carried out by APOBEC enzymes. Both types of RNA editing can modify RNA recognition motifs, altering the interaction and regulatory roles of RBPs with a certain degree of sequence specificity (Lerner, Papavasiliou, and Pecori 2018). Here, we were interested in evaluating the impact of C-to-U editing in the recognition of DRACH motifs and subsequent m^{6}A deposition. We thus aimed to assess the potential interplay

between the APOBEC1-mediated C-to-U activity and m^6A modification through transcriptome-wide analysis using nanopore sequencing data.

For the study of RNA-editing, samples derived from a HEK293T cell line, previously induced with APOBEC1 and its cofactor RBM47, were analysed, while the WT (non-induced samples) served as negative controls. The mapping of edited positions was performed using REDItools software (Lo Giudice et al. 2020). In parallel, the m6anet tool was applied for the detection of m^6A sites throughout the transcriptome (**Figure 4.1A**) (Hendra et al. 2022). repeated elements, single nucleotide polymorphisms (SNPs), genes names and additional features were annotated with the annotatedTable.py and commonSNPs.py scripts from REDItools. Then repeated elements and common SNPs were dropped. To keep track of only real editing sites exclusively a second filtering step was applied in which only those sites with a substitution frequency > 0.05 and < 0.5 were selected. The explanation for this filtering criteria is that RNA editing are relatively unusual events compared with genomic variants (SNPs), so that sites with substitution frequency higher than 0.5 can be considered as SNPs rather than RNA editing.

Nanopore reads have an overall lower quality than Illumina reads. The coverage obtained for each position was also lower in nanopore. Considering that most of the median coverage is around 20-30 reads, only those sites with a coverage higher than 10 reads were retained, being considerably less restrictive than Illumina standard thresholds (approximately 30-50 reads). Editing sites where the substitution was not frequent enough were also discarded using a base supported variation of at least 3 times as cutoff value.

In a preliminary analysis, a considerable number of C-to-U editing events was observed in the WT samples i.e. the samples without hA1 induced editing. At this point, it is worth mentioning that nanopore technology can introduce C-to-U errors on the sequences, raising false positive editing events. In addition, nanopore has a high error rate and this negatively impacts the read quality. For this reason, an additional step of false positive correction was necessary to assess real editing. More precisely, we performed a second round of selection of real C-to-U editing using iForest (Fonzino et al. 2024), a predictive tool designed to discard nanopore errors and keep only real editing events. However, despite the filtering steps, nanopore samples still contained too much noise (**Figure 4.1B**). In addition, the number of editing sites obtained after the filtering was very low if we compare it with those obtained with Illumina sequencing. This could also explain why we observed such a small overlap between m^6A and C-to-U modified transcripts (**Figure 4.1C**). Since the RNA-editing sites detected with nanopore did not resulted reliable enough, Illumina sequencing was

additionally considered for detection of C-to-U events and subsequent match with m^6A as an alternative approach.

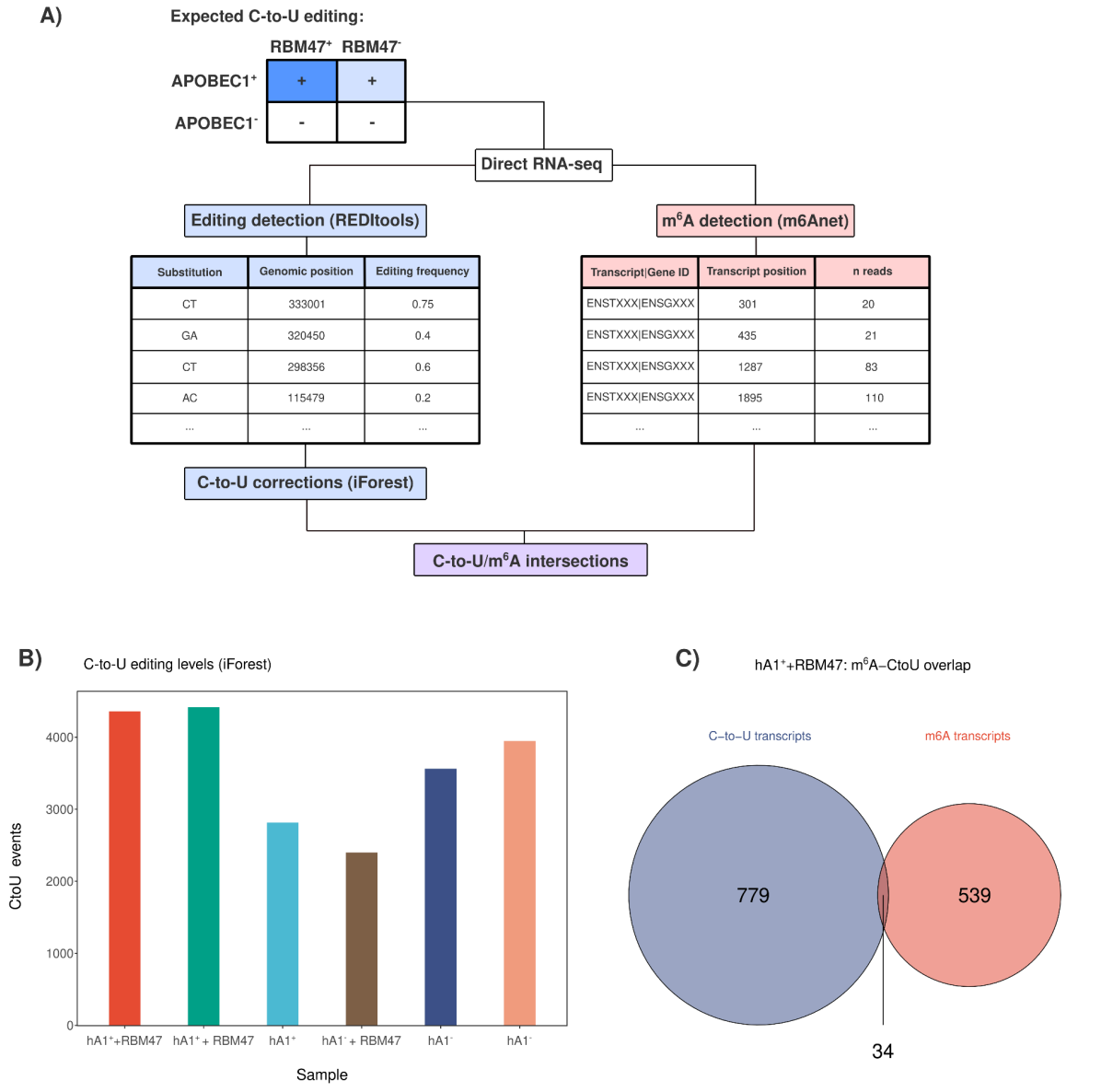


Figure 4.1: Match between C-to-U editing and m^6A using nanopore direct sequencing data *(**A**) Graphical abstract. Steps followed for preprocessing of nanopore data and detection of C-o-U and m^6A. (**B**) Barplots showing the C-to-U editing levels per sample after conducting iForest for false positive removal. (**C**) Euler plot displaying the overlap between transcripts with reported C-to-U and m^6A in a $hA1^+$+ RMB47 sample.*

m^6A deposition promotes translation of DNA methyltransferase 1 contributing to progression of ovarian cancer.

(Kumari et al., in collaboration with the laboratory of Prof. Francesca Aguilo)

DNA methyltransferase 1 (DNMT1) plays a critical role in tumor progression by repressing the transcription of tumor suppressor genes through DNA methylation thereby contributing to poorer prognostic outcomes. Methylated RNA immunoprecipitation sequencing (MeRIP-seq) analysis of OVCAR8 cells which represents high grade serous ovarian cancer, demonstrated that CDS of DNMT1 is modified by m^6A. The ongoing work of Kumari et al. demonstrates that lentiviral-mediated knockdown of METTL3 resulted into reduced DNMT1 protein level without affecting the levels of DNMT1 transcript, indicating the role of METTL3-mediated m^6A in DNMT1 protein translation and/ or stability.

High-grade serous ovarian cancer (HGSOC) is distinguished by a highly complex tumor microenvironment. Tumours comprise diverse cellular populations, including immune cells, fibroblasts, endothelial cells, and epithelial cells. Understanding the gene expression profiles within these distinct cell types is crucial for capturing a comprehensive picture of tumor biology. In this study, we aimed to investigate the expression patterns of m^6A modifiers across different cell types in HGSOC tumours and compare them to those in normal ovarian tissues. To assess this, we performed statistical analysis over a pre-existing single-cell RNA sequencing (scRNA-Seq) dataset (GSE184880) that included seven samples from HGSOC patients and five normal ovarian samples (Xu et al. 2022).

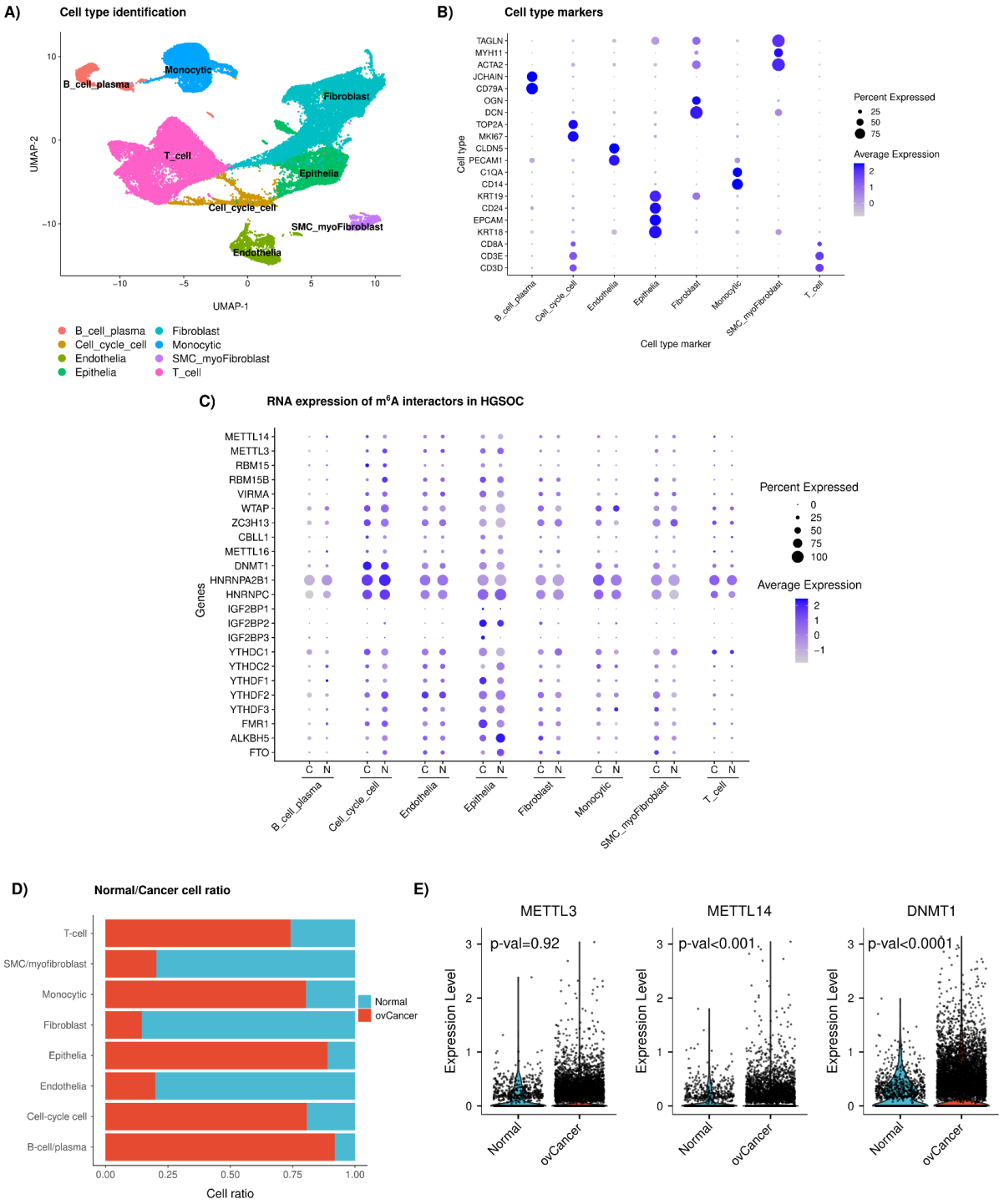


Figure 4.2: m^{6}A regulators and DMNT1 expression across cell types in HGSOC: *(**A**) UMAP projection of integrated HGSOC and normal samples. Each point represents a single cell and the color code is assigned according to the cell type. (**B**) Bubble plot of molecular markers used as criteria for cell type assignment. Each column on the x axis represents a different cell type and each row on the y axis is a molecular marker. The bubble size and the color intensity represents respectively the percentage of cells expressing the marker and the average expression (**C**) Cancer vs normal bubble plot for expression of DNMT1 and m^{6}A*

*regulators. Each cell type is splitted into two columns representing the expression in cancer and normal cells (C and N) (**D**) Stacked barplots of cell ratio. The x axis represents the ratio of cells and each row in the y axis is a different cell type. Blue bars represent the fraction of normal cells and red bars the fraction of cancer cells. (**E**) Jitter boxplot sets for gene expression levels of METTL3 METTL14 and DNMT1 in epithelial cells. The displayed p-values correspond to the t-test for independent samples adjusted using the BH method.*

Materials & methods

Single-cell dataset is available at the GEO accession series GSE184880 (Xu et al. 2022). Downstream analysis was conducted with Seurat R package (v5.1.0) and scCustomize (v2.1.2) was used to improve feature visualization (Hao et al. 2024; Marsh, Salmon, and Hoffman 2024). Each sample was preprocessed according to the protocol described in (Xu et al. 2022). The preprocessing steps included an initial quality control where cells with fewer than 200 features and mitochondrial percent higher than 40% were filtered out thus obtaining a subset of 59,664 isolated cells.

Sample datasets were then merged into a combined object. RNA expression data were normalized using the log-normalization method and scaled to avoid batch effect bias. We thus proceeded with the standard single-cell workflow to obtain an overview of the combined data. We performed Principal component analysis (PCA) for dimensionality reduction followed by Louvain clustering. The optimal number of clusters was estimated using the elbow method. The elbow was observed approximately between 8-10 clusters, being consistent with the number of cell types (8) originally identified in (Xu et al. 2022). Resolution was adjusted according to the expected number of clusters. Finally we visualized the two dimensional scatter plots using the Uniform manifold approximation and projection (UMAP). Data was subsequently integrated including anchor-based correction to remove batch effect. The integration step was followed by a second round of clustering and UMAP. Finally, the cell type identity was assigned for each cluster according to the gene markers - cell type associations described in (Xu et al. 2022).

After cell type assignment, we redirected our analysis towards DNMT1 and the m^6A writers METTL3, METTL14, and CBLL1, along with other well-known m^6A interactors. Differentially expressed genes were identified using the FindMarkers() function from the Seurat package. For each cell type t-test comparison of RNA expression levels were performed between cells

from cancer and normal samples, with p-values adjusted using the Benjamini-Hochberg method.

Conclusions

In this thesis, I have explored the complexity of post-transcriptional regulatory networks and addressed the pivotal roles of RNA-binding proteins (RBPs) in such regulation. More precisely, my work has focused on the regulatory interplay between RBPs which can give rise to multiple RBP-RBP interactions.

Although RBP-RBP interactions are characterized by their context dependency, our findings suggest the existence of general tendencies toward overlapping or nearby RNA-binding sites, which could be indicative of competitive and cooperative behaviours. The results concerning the NMF clusters, although inconclusive, point to the possibility of predicting cooperation and competition between RBPs in the future. In this context, the integration of additional features such as transcript regions, RNA secondary structures, and protein architecture information could considerably enrich the analysis and shed light on the intricate mechanisms of RBP-RBP interactions.

To summarize, the findings presented in this thesis emphasize the existence of cooperative and competitive RBP mechanisms and the influence of PTMs in such interactions. Specifically, we showed that the stratification of intersecting and proximal RNA-binding sites, while still insufficient, can facilitate the discrimination between cooperation and competition.

Having addressed competitive and cooperative RBP-RBP interactions, we shifted the focus of this work to post-translational modifications (PTMs) of these RBPs, a key element that has been shown to be determinant for such interplays. In this regard, we discovered that the presence of acetylation and methylation sites on RBPs subdivided the pool of RBP pairs in different groups having significant differences in features that are descriptive of RBP-RBP interactions, thereby suggesting a yet unappreciated contribution of these PTMs to cooperative and competitive behaviors.

In addition, while acetylation resulted in a negative impact over the RNA-binding activity of RBPs, our analysis also revealed that acetylation sites increased the probability of RBP-RBP interactions. These results combined suggest that acetylation potentially induces shifts in RBP behavior, decreasing the affinity for target RNAs and facilitating PPIs. Such shifts triggered by acetylation could be explained by structural changes within the RBP that leads to conformations of RBDs which are less likely to interact with RNA molecules, and to PPI domains being more receptive.

In this work, we also have shown that the frequencies of intersecting and nearby binding sites can be predicted based on the presence of potential methylation and acetylation sites and vice versa. Furthermore, we evaluated the contribution of each of the features to the obtained predictions, revealing that the distance between proximal RNA-binding sites, the length of intersecting RNA-binding sites, and motif similarity were among the most informative predictors of presence of PTM sites. Summarizing, the presence of PTM sites was proven to be definitory of differential RBP-RBP interactions, even though their specific impact on these interactions is still unclear.

Parallel to these findings, we also conducted an analysis of RNA-Seq data derived from breast cancer samples to assess differential expression of RBPs involved in RNA modifications. The differential expression analysis for the SCAN-B cohort of patients revealed 8 RMPs overexpressed in the triple negative subtype. While the association between the interactors of each RNA modification has been the focus of the scientific community for the last two decades, the crosstalk between RMPs of different RNA modifications has not been so thoroughly explored yet.

This work emphasizes the need to address potential crosstalks between RMPs mediated by multiple RNA modifications. For instance, the overexpression of APOBECs observed in the TNBC samples indicates a potential APOBEC-mediated C-to-U editing activity which might interfere with the deposition of m^6A and m^5C. Moreover, it still remains unclear if RNA methylation can impact on the RNA editing efficiency. In this context, the interplay between these factors could shed light on how the disruption of several post-transcriptional processes may lead to the development and progression of cancer and other diseases.

The m^6A readers IGF2BP2 and IGF2BP3 were among the RMPs overexpressed in the TNBC. Considering the role of IGF2BPs as enhancers of mRNA stability, our findings suggest that bread cancer development, and particularly the TNBC subtype, could be associated with an increased stability of m^6A-marked mRNAs, which includes several

oncogenic transcripts such as c-MYC and KRAS. Although the upregulation of the IGF2BPs might suggest m^6A-mediated stabilization of transcripts, this does not necessarily indicate a global effect across the entire transcriptome. Such upregulation may instead reflect more specific regulatory events targeting a subset of m^6A modified RNAs. Furthermore, the present work does not evaluate the influence of additional factors independent of m^6A, which could also play critical roles in mRNA stabilization.

Acknowledgements

First and foremost, I would like to express my gratitude to Prof. Erik Dassi for his mentorship and guidance throughout the years of my Ph.D. journey. Your expertise in bioinformatics and insightful feedback have been invaluable in shaping my academic growth and the development of this thesis. Thanks for granting me the opportunity to develop my doctoral program within the laboratory of RNA Regulatory Networks.

In addition, I am deeply grateful to Prof Francesca Aguilo and Prof Sivo Conticello who deserve a special mention for their pivotal role as supervisors during my secondments, guiding me through new areas of knowledge. Your guidance, patience and dedication not only helped me acquire essential skills but also inspired me to approach challenges with curiosity and determination. The insightful discussions I have had with both of you have been instrumental in enriching my research experience and personal development.

I would like to extend my gratitude to Prof. Alessandro Romanel and Prof. Alessandro Quattrone for their perceptive advice and reviews of my work as members of the annual CIBIO evaluation committee. Your constructive feedback during the evaluations have played a key role in refining my analyses and ensuring the research progression. I deeply appreciate your honesty and objective suggestions which have significantly contributed to the successful completion of this thesis.

Finally, I would also like to express my gratitude to all the PIs and ESRs from the ROPES consortium and Lorena Cebolla, Thank you all for your time dedicated to the organization of the training schools, which have been enriching experiences for my research career and, of course, thank you all for the great times we spent together.

References

Acera Mateos, P., A. J Sethi, A. Ravindran, A. Srivastava, K. Woodward, S. Mahmud, M. Kanchi, et al. 2024. “Prediction of m6A and m5C at Single-Molecule Resolution Reveals a Transcriptome-Wide Co-Occurrence of RNA Modifications.” *Nature Communications* 15 (1): 3899.

Alwarawrah, Yazan, Philip Hughes, David Loiselle, David A. Carlson, David B. Darr, Jamie L. Jordan, Jessie Xiong, et al. 2016. “Fasnall, a Selective FASN Inhibitor, Shows Potent Anti-Tumor Activity in the MMTV-Neu Model of HER2(+) Breast Cancer.” *Cell Chemical Biology* 23 (6): 678–88.

Archuleta, Stephen R., James A. Goodrich, and Jennifer F. Kugel. 2024. “Mechanisms and Functions of the RNA Polymerase II General Transcription Machinery during the Transcription Cycle.” *Biomolecules* 14 (2): 176.

Arnesen, Thomas. 2011. “Towards a Functional Understanding of Protein N-Terminal Acetylation.” *PLoS Biology* 9 (5): e1001074.

Bastide, Amandine, and Alexandre David. 2018. “Interaction of rRNA with mRNA and tRNA in Translating Mammalian Ribosome: Functional Implications in Health and Disease.” *Biomolecules* 8 (4): 100.

Bawankar, Praveen, Tina Lence, Chiara Paolantoni, Irmgard U. Haussmann, Migle Kazlauskiene, Dominik Jacob, Jan B. Heidelberger, et al. 2021. “Hakai Is Required for Stabilization of Core Components of the m6A mRNA Methylation Machinery.” *Nature Communications* 12 (1): 3778.

Berenson, Anna, Ryan Lane, Luis F. Soto-Ugaldi, Mahir Patel, Cosmin Ciausu, Zhaorong Li, Yilin Chen, et al. 2023. “Paired Yeast One-Hybrid Assays to Detect DNA-Binding Cooperativity and Antagonism across Transcription Factors.” *Nature Communications* 14 (1): 6570.

Bianconi, Eva, Allison Piovesan, Federica Facchin, Alina Beraudi, Raffaella Casadei, Flavia Frabetti, Lorenza Vitale, et al. 2013. “An Estimation of the Number of Cells in the Human Body.” *Annals of Human Biology* 40 (6): 463–71.

Bohnsack, Katherine E., Claudia Höbartner, and Markus T. Bohnsack. 2019. “Eukaryotic 5-Methylcytosine (m^5C) RNA Methyltransferases: Mechanisms, Cellular Functions, and Links to Disease.” *Genes* 10 (2): 102.

Bourdon, Sébastien, Pauline Herviou, Leïla Dumas, Eliana Destefanis, Andrea Zen, Anne Cammas, Stefania Millevoi, and Erik Dassi. 2023. “QUADRatlas: The RNA G-Quadruplex and RG4-Binding Proteins Database.” *Nucleic Acids Research* 51 (D1): D240–47.

Canisius, Sander, John W. M. Martens, and Lodewyk F. A. Wessels. 2016. “A Novel Independence Test for Somatic Alterations in Cancer Shows That Biology Drives Mutual Exclusivity but Chance Explains Most Co-Occurrence.” *Genome Biology* 17 (1): 261.

Castello, Alfredo, Bernd Fischer, Christian K. Frese, Rastislav Horos, Anne-Marie Alleaume, Sophia Foehr, Tomaz Curk, Jeroen Krijgsveld, and Matthias W. Hentze. 2016. “Comprehensive Identification of RNA-Binding Domains in Human Cells.” *Molecular Cell* 63 (4): 696–710.

Catalanotto, Caterina, Christian Barbato, Carlo Cogoni, and Dario Benelli. 2023. “The RNA-Binding Function of Ribosomal Proteins and Ribosome Biogenesis Factors in Human Health and Disease.” *Biomedicines* 11 (11): 2969.

Chen, C. Y., and A. B. Shyu. 1995. “AU-Rich Elements: Characterization and Importance in mRNA Degradation.” *Trends in Biochemical Sciences* 20 (11): 465–70.

Chen, Tianqi, and Carlos Guestrin. 2016. “XGBoost: A Scalable Tree Boosting System.” *arXiv [cs.LG]*. arXiv. https://github.com/dmlc/xgboost.

Cockman, Eric, Paul Anderson, and Pavel Ivanov. 2020. “TOP mRNPs: Molecular Mechanisms and Principles of Regulation.” *Biomolecules* 10 (7). https://doi.org/10.3390/biom10070969.

Corbett, Anita H. 2018. “Post-Transcriptional Regulation of Gene Expression and Human

Disease.” *Current Opinion in Cell Biology* 52 (June):96–104.
Corley, Meredith, Margaret C. Burns, and Gene W. Yeo. 2020. “How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms.” *Molecular Cell* 78 (1): 9–29.
Cruciani, Sonia, Anna Delgado-Tejedor, Leszek P. Pryszcz, Rebeca Medina, Laia Llovera, and Eva Maria Novoa. 2023. “De Novo Basecalling of m6A Modifications at Single Molecule and Single Nucleotide Resolution.” *bioRxiv*. https://doi.org/10.1101/2023.11.13.566801.
Csárdi, Gábor, and T. Nepusz. 2006. “The Igraph Software Package for Complex Network Research” 1695. https://static1.squarespace.com/static/5b68a4e4a2772c2a206180a1/t/5cd1e3cbb208fc26c99de080/1557259212150/c1602a3c126ba822d0bc4293371c.pdf.
Cui, Li, Yucheng Zheng, Rongwei Xu, Yunfan Lin, Jiarong Zheng, Pei Lin, Bing Guo, Shuyu Sun, and Xinyuan Zhao. 2024. “Alternative Pre-mRNA Splicing in Stem Cell Function and Therapeutic Potential: A Critical Review of Current Evidence.” *International Journal of Biological Macromolecules* 268 (Pt 1): 131781.
Curtis, Christina, Sohrab P. Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, et al. 2012. “The Genomic and Transcriptomic Architecture of 2,000 Breast Tumours Reveals Novel Subgroups.” *Nature* 486 (7403): 346–52.
Daitoku, Hiroaki, Jun-Ichi Sakamaki, and Akiyoshi Fukamizu. 2011. “Regulation of FoxO Transcription Factors by Acetylation and Protein-Protein Interactions.” *Biochimica et Biophysica Acta* 1813 (11): 1954–60.
Dai, Xiaofeng, Ting Li, Zhonghu Bai, Yankun Yang, Xiuxia Liu, Jinling Zhan, and Bozhi Shi. 2015. “Breast Cancer Intrinsic Subtype Classification, Clinical Use and Future Trends.” *American Journal of Cancer Research* 5 (10): 2929–43.
Dassi, Erik. 2017. “Handshakes and Fights: The Regulatory Interplay of RNA-Binding Proteins.” *Frontiers in Molecular Biosciences* 4 (September):67.
Destefanis, Eliana, Gülben Avşar, Paula Groza, Antonia Romitelli, Serena Torrini, Pınar Pir, Silvestro G. Conticello, Francesca Aguilo, and Erik Dassi. 2021. “A Mark of Disease: How mRNA Modifications Shape Genetic and Acquired Pathologies.” *RNA* 27 (4): 367–89.
Destefanis, Eliana, Denise Sighel, Davide Dalfovo, Riccardo Gilmozzi, Francesca Broso, Andrea Cappannini, Janusz M. Bujnicki, Alessandro Romanel, Erik Dassi, and Alessandro Quattrone. 2024. “The Three YTHDF Paralogs and VIRMA Are Strong Cross-Histotype Tumor Driver Candidates among m6A Core Genes.” *NAR Cancer* 6 (4): zcae040.
Driscoll, D. M., J. K. Wynne, S. C. Wallis, and J. Scott. 1989. “An in Vitro System for the Editing of Apolipoprotein B mRNA.” *Cell* 58 (3): 519–25.
Dumas, Leïla, Pauline Herviou, Erik Dassi, Anne Cammas, and Stefania Millevoi. 2021. “G-Quadruplexes in RNA Biology: Recent Advances and Future Directions.” *Trends in Biochemical Sciences* 46 (4): 270–83.
Egbert, C. M., L. R. Warr, K. L. Pennington, M. M. Thornton, A. J. Vaughan, S. W. Ashworth, M. J. Heaton, N. English, M. P. Torres, and J. L. Andersen. 2023. “The Integration of Proteome-Wide PTM Data with Protein Structural and Sequence Features Identifies Phosphorylations That Mediate 14-3-3 Interactions.” *Journal of Molecular Biology* 435 (2): 167890.
Elcheva, Irina A., Chethana P. Gowda, Daniel Bogush, Svetlana Gornostaeva, Anna Fakhardo, Neil Sheth, Kathleen M. Kokolus, et al. 2023. “IGF2BP Family of RNA-Binding Proteins Regulate Innate and Adaptive Immune Responses in Cancer Cells and Tumor Microenvironment.” *Frontiers in Immunology* 14 (July):1224516.
Esteva-Socias, Margalida, and Francesca Aguilo. 2024. “METTL3 as a Master Regulator of Translation in Cancer: Mechanisms and Implications.” *NAR Cancer* 6 (1): zcae009.
Fang, Yan, Xinyu Liu, Yuli Liu, and Naiyi Xu. 2024. “Insights into the Mode and Mechanism of Interactions between RNA and RNA-Binding Proteins.” *International Journal of Molecular Sciences* 25 (21): 11337.

Fang, Zheng, Hong-Yu Shen, Qi Xu, Hong-Lei Zhou, Lei Li, Si-Yuan Yang, Zhen Zhu, and Jin-Hai Tang. 2022. “PUS1 Is a Novel Biomarker for Predicting Poor Outcomes and Triple-Negative Status in Breast Cancer.” *Frontiers in Oncology* 12 (November):1030571.
Fernandes, Juliane C. R., Stephanie M. Acuña, Juliana I. Aoki, Lucile M. Floeter-Winter, and Sandra M. Muxel. 2019. “Long Non-Coding RNAs in the Regulation of Gene Expression: Physiology and Disease.” *Non-Coding RNA* 5 (1): 17.
Fierro-Monti, Ivo. 2024. “RBPs: An RNA Editor’s Choice.” *Frontiers in Molecular Biosciences* 11 (August):1454241.
Fonzino, Adriano, Caterina Manzari, Paola Spadavecchia, Uday Munagala, Serena Torrini, Silvestro Conticello, Graziano Pesole, and Ernesto Picardi. 2024. “Unraveling C-to-U RNA Editing Events from Direct RNA Sequencing.” *RNA Biology* 21 (1): 1–14.
Frankish, Adam, Mark Diekhans, Irwin Jungreis, Julien Lagarde, Jane E. Loveland, Jonathan M. Mudge, Cristina Sisu, et al. 2021. “GENCODE 2021.” *Nucleic Acids Research* 49 (D1): D916–23.
Fredericks, Alger M., Kamil J. Cygan, Brian A. Brown, and William G. Fairbrother. 2015. “RNA-Binding Proteins: Splicing Factors and Disease.” *Biomolecules* 5 (2): 893–909.
Frye, Michaela, Ilaria Dragoni, Suet-Feung Chin, Immaculada Spiteri, Agata Kurowski, Elena Provenzano, Andrew Green, et al. 2010. “Genomic Gain of 5p15 Leads to over-Expression of Misu (NSUN2) in Breast Cancer.” *Cancer Letters* 289 (1): 71–80.
Fumagalli, Debora, David Gacquer, Françoise Rothé, Anne Lefort, Frederick Libert, David Brown, Naima Kheddoumi, et al. 2015. “Principles Governing A-to-I RNA Editing in the Breast Cancer Transcriptome.” *Cell Reports* 13 (2): 277–89.
Fu, Xiang-Dong. 2014. “Non-Coding RNA: A New Frontier in Regulatory Biology.” *National Science Review* 1 (2): 190–204.
Gallardo-Dodd, Carlos J., and Claudia Kutter. 2024. “The Regulatory Landscape of Interacting RNA and Protein Pools in Cellular Homeostasis and Cancer.” *Human Genomics* 18 (1): 109.
Gannon, Hugh S., Tao Zou, Michael K. Kiessling, Galen F. Gao, Diana Cai, Peter S. Choi, Alexandru P. Ivan, et al. 2018. “Identification of ADAR1 Adenosine Deaminase Dependency in a Subset of Cancer Cells.” *Nature Communications* 9 (1): 5450.
Gao, Na, Yueheng Li, Jing Li, Zhengfan Gao, Zhenzhen Yang, Yong Li, Hongtao Liu, and Tianli Fan. 2020. “Long Non-Coding RNAs: The Regulatory Mechanisms, Research Strategies, and Future Directions in Cancers.” *Frontiers in Oncology* 10 (December):598817.
Gardiner, Amy S., Jeffery L. Twiss, and Nora I. Perrone-Bizzozero. 2015. “Competing Interactions of RNA-Binding Proteins, MicroRNAs, and Their Targets Control Neuronal Development and Function.” *Biomolecules* 5 (4): 2903–18.
Gassner, Franz J., Nadja Zaborsky, Ilana Buchumenski, Erez Y. Levanon, Matthias Gatterbauer, Maria Schubert, Stefanie Rauscher, et al. 2021. “RNA Editing Contributes to Epitranscriptome Diversity in Chronic Lymphocytic Leukemia.” *Leukemia* 35 (4): 1053–63.
Gaujoux, Renaud, and Cathal Seoighe. 2010. “A Flexible R Package for Nonnegative Matrix Factorization.” *BMC Bioinformatics* 11 (1): 367.
Geffen, Yifat, Shankara Anand, Yo Akiyama, Tomer M. Yaron, Yizhe Song, Jared L. Johnson, Akshay Govindan, et al. 2023. “Pan-Cancer Analysis of Post-Translational Modifications Reveals Shared Patterns of Protein Regulation.” *Cell* 186 (18): 3945–67.e26.
Ge, Junhui, and Yi-Tao Yu. 2013. “RNA Pseudouridylation: New Insights into an Old Modification.” *Trends in Biochemical Sciences* 38 (4): 210–18.
Gerstberger, Stefanie, Markus Hafner, and Thomas Tuschl. 2014. “A Census of Human RNA-Binding Proteins.” *Nature Reviews. Genetics* 15 (12): 829–45.
Glisovic, Tina, Jennifer L. Bachorik, Jeongsik Yong, and Gideon Dreyfuss. 2008. “RNA-Binding Proteins and Post-Transcriptional Gene Regulation.” *FEBS Letters* 582 (14): 1977–86.
Gott, J. M., and R. B. Emeson. 2000. “Functions and Mechanisms of RNA Editing.” *Annual*

Review of Genetics 34 (1): 499–531.
Gu, Chang, Xin Shi, Chenyang Dai, Feng Shen, Gaetano Rocco, Jiafei Chen, Zhengyu Huang, et al. 2020. “RNA m6A Modification in Cancers: Molecular Mechanisms and Potential Clinical Applications.” *Innovation (Cambridge (Mass.))* 1 (3): 100066.
Hafner, Markus, Maria Katsantoni, Tino Köster, James Marks, Joyita Mukherjee, Dorothee Staiger, Jernej Ule, and Mihaela Zavolan. 2021. “CLIP and Complementary Methods.” *Nature Reviews Methods Primers* 1 (1): 1–23.
Hao, Yuhan, Tim Stuart, Madeline H. Kowalski, Saket Choudhary, Paul Hoffman, Austin Hartman, Avi Srivastava, et al. 2024. “Dictionary Learning for Integrative, Multimodal and Scalable Single-Cell Analysis.” *Nature Biotechnology* 42 (2): 293–304.
Hendra, Christopher, Ploy N. Pratanwanich, Yuk Kei Wan, W. S. Sho Goh, Alexandre Thiery, and Jonathan Göke. 2022. “Detection of m6A from Direct RNA Sequencing Using a Multiple Instance Learning Framework.” *Nature Methods* 19 (12): 1590–98.
Hentze, Matthias W., Alfredo Castello, Thomas Schwarzl, and Thomas Preiss. 2018. “A Brave New World of RNA-Binding Proteins.” *Nature Reviews. Molecular Cell Biology* 19 (5): 327–41.
Huang, Huilin, Hengyou Weng, Wenju Sun, Xi Qin, Hailing Shi, Huizhe Wu, Boxuan Simen Zhao, et al. 2018. “Recognition of RNA N6-Methyladenosine by IGF2BP Proteins Enhances mRNA Stability and Translation.” *Nature Cell Biology* 20 (3): 285–95.
Huang, Zheng-Hao, Yu-Ping Du, Jing-Tao Wen, Bing-Feng Lu, and Yang Zhao. 2022. “snoRNAs: Functions and Mechanisms in Biological Processes, and Roles in Tumor Pathophysiology.” *Cell Death Discovery* 8 (1): 259.
Hu, Lulu, Shun Liu, Yong Peng, Ruiqi Ge, Rui Su, Chamara Senevirathne, Bryan T. Harada, et al. 2022. “m6A RNA Modifications Are Measured at Single-Base Resolution across the Mammalian Transcriptome.” *Nature Biotechnology* 40 (8): 1210–19.
Hu, Yan-Xia, Li-Ting Diao, Ya-Rui Hou, Guo Lv, Shuang Tao, Wan-Yi Xu, Shu-Juan Xie, Ya-Han Ren, and Zhen-Dong Xiao. 2024. “Pseudouridine Synthase 1 Promotes Hepatocellular Carcinoma through mRNA Pseudouridylation to Enhance the Translation of Oncogenic mRNAs.” *Hepatology (Baltimore, Md.)* 80 (5): 1058–73.
Hwang, Woochang, Stefano Calza, Marco Silvestri, Yudi Pawitan, and Youngjo Lee. 2019. “CREDO: Highly Confident Disease-Relevant A-to-I RNA-Editing Discovery in Breast Cancer.” *Scientific Reports* 9 (1): 5064.
Ito, K., and D. Murphy. 2013. “Application of ggplot2 to Pharmacometric Graphics.” *CPT: Pharmacometrics & Systems Pharmacology* 2 (10): e79.
Ivanov, Gleb S., Tatyana Ivanova, Julia Kurash, Alexey Ivanov, Sergey Chuikov, Farid Gizatullin, Enrique M. Herrera-Medina, Frank Rauscher 3rd, Danny Reinberg, and Nickolai A. Barlev. 2007. “Methylation-Acetylation Interplay Activates p53 in Response to DNA Damage.” *Molecular and Cellular Biology* 27 (19): 6756–69.
Jalili, Pégah, Danae Bowen, Adam Langenbucher, Shinho Park, Kevin Aguirre, Ryan B. Corcoran, Angela G. Fleischman, Michael S. Lawrence, Lee Zou, and Rémi Buisson. 2020. “Quantification of Ongoing APOBEC3A Activity in Tumor Cells by Monitoring RNA Editing at Hotspots.” *Nature Communications* 11 (1): 2971.
Jin, Qin, Gang Liu, Biao Wang, Shubin Li, Kan Ni, Chunyu Wang, Jingyu Ren, Shu Zhang, and Yanfeng Dai. 2020. “High Methionyl-tRNA Synthetase Expression Predicts Poor Prognosis in Patients with Breast Cancer.” *Journal of Clinical Pathology* 73 (12): 803–12.
Jumper, John, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, et al. 2021. “Highly Accurate Protein Structure Prediction with AlphaFold.” *Nature* 596 (7873): 583–89.
June, Terry Therneau. n.d. “A Package for Survival Analysis in R.” https://cran.r-project.org/web/packages/survival/vignettes/survival.pdf.
Kawaf, Rawan R., Wafaa S. Ramadan, and Raafat El-Awady. 2024. “Deciphering the Interplay of Histone Post-Translational Modifications in Cancer: Co-Targeting Histone Modulators for Precision Therapy.” *Life Sciences* 346 (June):122639.
Khoroshkin, Matvei, Daniel Asarnow, Shaopu Zhou, Albertas Navickas, Aidan Winters,

Jackson Goudreau, Simon K. Zhou, et al. 2024. “A Systematic Search for RNA Structural Switches across the Human Transcriptome.” *Nature Methods* 21 (9): 1634–45.
Khoroshkin, Matvei, Andrey Buyan, Martin Dodel, Albertas Navickas, Johnny Yu, Fathima Trejo, Anthony Doty, et al. 2023. “Systematic Identification of Post-Transcriptional Regulatory Modules.” *bioRxiv*. https://doi.org/10.1101/2023.02.27.530345.
Kim, Eun Young, Ji Ye Jung, Arum Kim, Kwangsoo Kim, and Yoon Soo Chang. 2017. “Methionyl-tRNA Synthetase Overexpression Is Associated with Poor Clinical Outcomes in Non-Small Cell Lung Cancer.” *BMC Cancer* 17 (1): 467.
Kiss, Arnold M., Beáta E. Jády, Edouard Bertrand, and Tamás Kiss. 2004. “Human Box H/ACA Pseudouridylation Guide RNA Machinery.” *Molecular and Cellular Biology* 24 (13): 5797–5807.
Kumari, Kanchan, Paula Groza, and Francesca Aguilo. 2021. “Regulatory Roles of RNA Modifications in Breast Cancer.” *NAR Cancer* 3 (3): zcab036.
Lachiondo-Ortega, Sofia, Claudia M. Rejano-Gordillo, Jorge Simon, Fernando Lopitz-Otsoa, Teresa C Delgado, Krystyna Mazan-Mamczarz, Naroa Goikoetxea-Usandizaga, et al. 2024. “SUMOylation Controls Hu Antigen R Posttranscriptional Activity in Liver Cancer.” *Cell Reports* 43 (3): 113924.
Lawrence, Michael, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T. Morgan, and Vincent J. Carey. 2013. “Software for Computing and Annotating Genomic Ranges.” *PLoS Computational Biology* 9 (8): e1003118.
Lee, Abigail V., Kevin A. Nestler, and Katherine B. Chiappinelli. 2024. “Therapeutic Targeting of DNA Methylation Alterations in Cancer.” *Pharmacology & Therapeutics* 258 (108640): 108640.
Lee, Flora C. Y., and Jernej Ule. 2018. “Advances in CLIP Technologies for Studies of Protein-RNA Interactions.” *Molecular Cell* 69 (3): 354–69.
Lee, Stuart, Dianne Cook, and Michael Lawrence. 2019. “Plyranges: A Grammar of Genomic Data Transformation.” *Genome Biology* 20 (1): 4.
Lehmann, Brian D., Joshua A. Bauer, Xi Chen, Melinda E. Sanders, A. Bapsi Chakravarthy, Yu Shyr, and Jennifer A. Pietenpol. 2011. “Identification of Human Triple-Negative Breast Cancer Subtypes and Preclinical Models for Selection of Targeted Therapies.” *The Journal of Clinical Investigation* 121 (7): 2750–67.
Lerner, Taga, F. Nina Papavasiliou, and Riccardo Pecori. 2018. “RNA Editors, Cofactors, and mRNA Targets: An Overview of the C-to-U RNA Editing Machinery and Its Implication in Human Disease.” *Genes* 10 (1): 13.
Levi, Ofri, and Yoav S. Arava. 2021. “Pseudouridine-Mediated Translation Control of mRNA by Methionine Aminoacyl tRNA Synthetase.” *Nucleic Acids Research* 49 (1): 432–43.
Li, Dandan, Yanhui Yang, Ruifang Li, Liang Huang, Zichao Wang, Qiwu Deng, and Shuaibo Dong. 2021. “N-Terminal Acetylation of Antimicrobial Peptide L163 Improves Its Stability against Protease Degradation.” *Journal of Peptide Science: An Official Publication of the European Peptide Society* 27 (9): e3337.
Li, Die, Shiqi Hu, Jiarong Ye, Chaojie Zhai, Jipeng Liu, Zuao Wang, Xinchi Zhou, Leifeng Chen, and Fan Zhou. 2024. “The Emerging Role of IGF2BP2 in Cancer Therapy Resistance: From Molecular Mechanism to Future Potential.” *International Journal of Molecular Sciences* 25 (22): 12150.
Li, Lin, Chongying Zhu, Shouying Xu, Qiang Xu, Da Xu, Sishun Gan, Xingang Cui, and Chao Tang. 2023. “PUS1 Is a Novel Biomarker for Evaluating Malignancy of Human Renal Cell Carcinoma.” *Aging* 15 (11): 5215–27.
Liu, Jianzhao, Yanan Yue, Dali Han, Xiao Wang, Ye Fu, Liang Zhang, Guifang Jia, et al. 2014. “A METTL3-METTL14 Complex Mediates Mammalian Nuclear RNA N6-Adenosine Methylation.” *Nature Chemical Biology* 10 (2): 93–95.
Liu, Ying, Ruihui Li, Jiawei Luo, and Zhaolei Zhang. 2022. “Inferring RNA-Binding Protein Target Preferences Using Adversarial Domain Adaptation.” *PLoS Computational Biology* 18 (2): e1009863.
Lo Giudice, Claudio, Marco Antonio Tangaro, Graziano Pesole, and Ernesto Picardi. 2020.

"Investigating RNA Editing in Deep Transcriptome Datasets with REDItools and REDIportal." *Nature Protocols* 15 (3): 1098–1131.
Long, Yunxin, Weng Hong Sou, Kristen Wing Yu Yung, Haizhen Liu, Stephanie Winn Chee Wan, Qingyun Li, Chuyue Zeng, et al. 2019. "Distinct Mechanisms Govern the Phosphorylation of Different SR Protein Splicing Factors." *The Journal of Biological Chemistry* 294 (4): 1312–27.
Love, Michael I., Wolfgang Huber, and Simon Anders. 2014. "Moderated Estimation of Fold Change and Dispersion for RNA-Seq Data with DESeq2." *Genome Biology* 15 (12): 550.
Luo, Yunhai, Benjamin C. Hitz, Idan Gabdank, Jason A. Hilton, Meenakshi S. Kagda, Bonita Lam, Zachary Myers, et al. 2020. "New Developments on the Encyclopedia of DNA Elements (ENCODE) Data Portal." *Nucleic Acids Research* 48 (D1): D882–89.
Lutz, Helmut, Vance Jaeger, Mischa Bonn, Jim Pfaendtner, and Tobias Weidner. 2017. "Acetylation Dictates the Morphology of Nanophase Biosilica Precipitated by a 14-Amino Acid Leucine-Lysine Peptide." *Journal of Peptide Science: An Official Publication of the European Peptide Society* 23 (2): 141–47.
Mancarella, Caterina, and Katia Scotlandi. 2019. "IGF2BP3 from Physiology to Cancer: Novel Discoveries, Unsolved Issues, and Future Perspectives." *Frontiers in Cell and Developmental Biology* 7:363.
Marsh, Samuel, Maëlle Salmon, and Paul Hoffman. 2024. *Samuel-marsh/scCustomize: Version 2.1.2*. Zenodo. https://doi.org/10.5281/ZENODO.10724532.
Martinez, Nicole M., Amanda Su, Margaret C. Burns, Julia K. Nussbacher, Cassandra Schaening, Shashank Sathe, Gene W. Yeo, and Wendy V. Gilbert. 2022. "Pseudouridine Synthases Modify Human Pre-mRNA Co-Transcriptionally and Affect Pre-mRNA Processing." *Molecular Cell* 82 (3): 645–59.e9.
Mayr, Christine. 2019. "What Are 3' UTRs Doing?" *Cold Spring Harbor Perspectives in Biology* 11 (10): a034728.
Mazzarello, P. 1999. "A Unifying Concept: The History of Cell Theory." *Nature Cell Biology* 1 (1): E13–15.
Micaelli, Mariachiara, Andrea Dalle Vedove, Linda Cerofolini, Jacopo Vigna, Denise Sighel, Sara Zaccara, Isabelle Bonomo, et al. 2022. "Small-Molecule Ebselen Binds to YTHDF Proteins Interfering with the Recognition of N 6-Methyladenosine-Modified RNAs." *ACS Pharmacology & Translational Science* 5 (10): 872–91.
Mignone, Flavio, Carmela Gissi, Sabino Liuni, and Graziano Pesole. 2002. "Untranslated Regions of mRNAs." *Genome Biology* 3 (3): REVIEWS0004.
Mina, Marco, Arvind Iyer, Daniele Tavernari, Franck Raynaud, and Giovanni Ciriello. 2020. "Discovering Functional Evolutionary Dependencies in Human Cancers." *Nature Genetics* 52 (11): 1198–1207.
Nag, Sharanya, Binita Goswami, Sukhen Das Mandal, and Partho Sarothi Ray. 2022. "Cooperation and Competition by RNA-Binding Proteins in Cancer." *Seminars in Cancer Biology* 86 (Pt 3): 286–97.
Naro, Chiara, Veronica Ruta, and Claudio Sette. 2024. "Splicing Dysregulation: Hallmark and Therapeutic Opportunity in Pancreatic Cancer." *Trends in Molecular Medicine*, December. https://doi.org/10.1016/j.molmed.2024.11.007.
Navarin, Nicolò, and Fabrizio Costa. 2017. "An Efficient Graph Kernel Method for Non-Coding RNA Functional Prediction." *Bioinformatics* 33 (17): 2642–50.
Nombela, Paz, Borja Miguel-López, and Sandra Blanco. 2021. "The Role of m6A, m5C and Ψ RNA Modifications in Cancer: Novel Therapeutic Opportunities." *Molecular Cancer* 20 (1): 18.
Orrantia-Borunda, Erasmo, Patricia Anchondo-Nuñez, Lucero Evelia Acuña-Aguilar, Francisco Octavio Gómez-Valles, and Claudia Adriana Ramírez-Valdespino. 2022. "Subtypes of Breast Cancer." In *Breast Cancer*, 31–42. Exon Publications.
Parker, Joel S., Michael Mullins, Maggie C. U. Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, et al. 2009. "Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes." *Journal of Clinical Oncology: Official Journal of the American*

Society of Clinical Oncology 27 (8): 1160–67.
Pecori, Riccardo, Salvatore Di Giorgio, J. Paulo Lorenzo, and F. Nina Papavasiliou. 2022. "Functions and Consequences of AID/APOBEC-Mediated DNA and RNA Deamination." *Nature Reviews. Genetics* 23 (8): 505–18.
Pereira, Bernard, Suet-Feung Chin, Oscar M. Rueda, Hans-Kristian Moen Vollan, Elena Provenzano, Helen A. Bardwell, Michelle Pugh, et al. 2016. "The Somatic Mutation Profiles of 2,433 Breast Cancers Refines Their Genomic and Transcriptomic Landscapes." *Nature Communications* 7 (May):11479.
Petljak, Mia, Alexandra Dananberg, Kevan Chu, Erik N. Bergstrom, Josefine Striepen, Patrick von Morgen, Yanyang Chen, et al. 2022. "Mechanisms of APOBEC3 Mutagenesis in Human Cancer Cells." *Nature* 607 (7920): 799–807.
Quattrone, Alessandro, and Erik Dassi. 2019. "The Architecture of the Human RNA-Binding Protein Regulatory Network." *iScience* 21 (November):706–19.
Quinlan, Aaron R., and Ira M. Hall. 2010. "BEDTools: A Flexible Suite of Utilities for Comparing Genomic Features." *Bioinformatics* 26 (6): 841–42.
Rainer, Johannes, Laurent Gatto, and Christian X. Weichenberger. 2019. "Ensembldb: An R Package to Create and Use Ensembl-Based Annotation Resources." *Bioinformatics* 35 (17): 3151–53.
Rauluseviciute, Ieva, Timothée Launay, Guido Barzaghi, Sarvesh Nikumbh, Boris Lenhard, Arnaud Regis Krebs, Jaime A. Castro-Mondragon, and Anthony Mathelier. 2023. "Identification of Transcription Factor Co-Binding Patterns with Non-Negative Matrix Factorization." *bioRxiv*. https://doi.org/10.1101/2023.04.28.538684.
Rueda, Oscar M., Stephen-John Sammut, Jose A. Seoane, Suet-Feung Chin, Jennifer L. Caswell-Jin, Maurizio Callari, Rajbir Batra, et al. 2019. "Dynamics of Breast-Cancer Relapse Reveal Late-Recurring ER-Positive Genomic Subgroups." *Nature* 567 (7748): 399–404.
Saal, Lao H., Johan Vallon-Christersson, Jari Häkkinen, Cecilia Hegardt, Dorthe Grabau, Christof Winter, Christian Brueffer, et al. 2015. "The Sweden Cancerome Analysis Network - Breast (SCAN-B) Initiative: A Large-Scale Multicenter Infrastructure towards Implementation of Breast Cancer Genomic Analyses in the Clinical Routine." *Genome Medicine* 7 (1): 20.
Salter, Jason D., Ryan P. Bennett, and Harold C. Smith. 2016. "The APOBEC Protein Family: United by Structure, Divergent in Function." *Trends in Biochemical Sciences* 41 (7): 578–94.
Schneider-Lunitz, Valentin, Jorge Ruiz-Orera, Norbert Hubner, and Sebastiaan van Heesch. 2021. "Multifunctional RNA-Binding Proteins Influence mRNA Abundance and Translational Efficiency of Distinct Sets of Target Genes." *PLoS Computational Biology* 17 (12): e1009658.
Schwartz, Schraga, Douglas A. Bernstein, Maxwell R. Mumbach, Marko Jovanovic, Rebecca H. Herbst, Brian X. León-Ricardo, Jesse M. Engreitz, et al. 2014. "Transcriptome-Wide Mapping Reveals Widespread Dynamic-Regulated Pseudouridylation of ncRNA and mRNA." *Cell* 159 (1): 148–62.
Seal, Ruth L., Bryony Braschi, Kristian Gray, Tamsin E. M. Jones, Susan Tweedie, Liora Haim-Vilmovsky, and Elspeth A. Bruford. 2023. "Genenames.org: The HGNC Resources in 2023." *Nucleic Acids Research* 51 (D1): D1003–9.
Sebestyén, Endre, Babita Singh, Belén Miñana, Amadís Pagès, Francesca Mateo, Miguel Angel Pujana, Juan Valcárcel, and Eduardo Eyras. 2016. "Large-Scale Analysis of Genome and Transcriptome Alterations in Multiple Tumors Unveils Novel Cancer-Relevant Splicing Networks." *Genome Research* 26 (6): 732–44.
Sender, Ron, and Ron Milo. 2021. "The Distribution of Cellular Turnover in the Human Body." *Nature Medicine* 27 (1): 45–48.
Shah, Ankeeta, Yingzhi Qian, Sebastien M. Weyn-Vanhentenryck, and Chaolin Zhang. 2017. "CLIP Tool Kit (CTK): A Flexible and Robust Pipeline to Analyze CLIP Sequencing Data." *Bioinformatics* 33 (4): 566–67.
Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel

Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. “Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks.” *Genome Research* 13 (11): 2498–2504.
“Spatial Correlation Statistics Enable Transcriptome-Wide Characterization of RNA Structure Binding.” 2021. *Cell Reports Methods* 1 (6): 100088.
Srikantan, Subramanya, Kumiko Tominaga, and Myriam Gorospe. 2012. “Functional Interplay between RNA-Binding Protein HuR and microRNAs.” *Current Protein & Peptide Science* 13 (4): 372–79.
Staaf, Johan, Jari Häkkinen, Cecilia Hegardt, Lao H. Saal, Siker Kimbung, Ingrid Hedenfalk, Tonje Lien, et al. 2022. “RNA Sequencing-Based Single Sample Predictors of Molecular Subtype and Risk of Recurrence for Clinical Assessment of Early-Stage Breast Cancer.” *NPJ Breast Cancer* 8 (1): 94.
Steele, Christopher D., Ammal Abbasi, S. M. Ashiqul Islam, Amy L. Bowes, Azhar Khandekar, Kerstin Haase, Shadi Hames-Fathi, et al. 2022. “Signatures of Copy Number Alterations in Human Cancer.” *Nature* 606 (7916): 984–91.
Sternburg, Erin L., Lara A. Gruijs da Silva, and Dorothee Dormann. 2022. “Post-Translational Modifications on RNA-Binding Proteins: Accelerators, Brakes, or Passengers in Neurodegeneration?” *Trends in Biochemical Sciences* 47 (1): 6–22.
Stockert, Jennifer A., Rachel Weil, Kamlesh K. Yadav, Natasha Kyprianou, and Ashutosh K. Tewari. 2021. “Pseudouridine as a Novel Biomarker in Prostate Cancer.” *Urologic Oncology* 39 (1): 63–71.
Street, Lena A., Katherine L. Rothamel, Kristopher W. Brannan, Wenhao Jin, Benjamin J. Bokor, Kevin Dong, Kevin Rhine, et al. 2024. “Large-Scale Map of RNA-Binding Protein Interactomes across the mRNA Life Cycle.” *Molecular Cell* 84 (19): 3790–3809.e8.
Street, Lena, Katherine Rothamel, Kristopher Brannan, Wenhao Jin, Benjamin Bokor, Kevin Dong, Kevin Rhine, et al. 2023. “Large-Scale Map of RNA Binding Protein Interactomes across the mRNA Life-Cycle.” *bioRxiv : The Preprint Server for Biology*, June. https://doi.org/10.1101/2023.06.08.544225.
Sun, Hanxiao, Kai Li, Cong Liu, and Chengqi Yi. 2023. “Regulation and Functions of Non-m6A mRNA Modifications.” *Nature Reviews. Molecular Cell Biology* 24 (10): 714–31.
Szklarczyk, Damian, Annika L. Gable, David Lyon, Alexander Junge, Stefan Wyder, Jaime Huerta-Cepas, Milan Simonovic, et al. 2019. “STRING v11: Protein-Protein Association Networks with Increased Coverage, Supporting Functional Discovery in Genome-Wide Experimental Datasets.” *Nucleic Acids Research* 47 (D1): D607–13.
Tai, Jiaojiao, Linbang Wang, Hao Guo, Ziqiang Yan, and Jingkun Liu. 2022. “Prognostic Implications of N6-Methyladenosine RNA Regulators in Breast Cancer.” *Scientific Reports* 12 (1): 1222.
Tang, Sze Jing, Haoqing Shen, Omer An, Huiqi Hong, Jia Li, Yangyang Song, Jian Han, et al. 2020. “Cis- and Trans-Regulations of Pre-mRNA Splicing by RNA Editing Enzymes Influence Cancer Development.” *Nature Communications* 11 (1): 799.
Taoka, Masato, Yuko Nobe, Yuka Yamaki, Ko Sato, Hideaki Ishikawa, Keiichi Izumikawa, Yoshio Yamauchi, et al. 2018. “Landscape of the Complete RNA Chemical Modifications in the Human 80S Ribosome.” *Nucleic Acids Research* 46 (18): 9289–98.
Thapar, Roopa. 2015. “Structural Basis for Regulation of RNA-Binding Proteins by Phosphorylation.” *ACS Chemical Biology* 10 (3): 652–66.
Thul, Peter J., Lovisa Åkesson, Mikaela Wiking, Diana Mahdessian, Aikaterini Geladaki, Hammou Ait Blal, Tove Alm, et al. 2017. “A Subcellular Map of the Human Proteome.” *Science* 356 (6340). https://doi.org/10.1126/science.aal3321.
“UniProt: The Universal Protein Knowledgebase in 2023.” 2023. *Nucleic Acids Research* 51 (D1): D523–31.
Vallon-Christersson, Johan. 2023. “RNA Sequencing-Based Single Sample Predictors of Molecular Subtype and Risk of Recurrence for Clinical Assessment of Early-Stage Breast Cancer.” Mendeley. https://doi.org/10.17632/YZXTXN4NMD.3.
Van Norden, Melissa, Zackary Falls, Sapan Mandloi, Brahm H. Segal, Bora E. Baysal, Ram

Samudrala, and Peter L. Elkin. 2024. “The Implications of APOBEC3-Mediated C-to-U RNA Editing for Human Disease.” *Communications Biology* 7 (1): 529.
Van Nostrand, Eric L., Peter Freese, Gabriel A. Pratt, Xiaofeng Wang, Xintao Wei, Rui Xiao, Steven M. Blue, et al. 2020. “A Large-Scale Binding and Functional Map of Human RNA-Binding Proteins.” *Nature* 583 (7818): 711–19.
Van Nostrand, Eric L., Gabriel A. Pratt, Alexander A. Shishkin, Chelsea Gelboin-Burkhart, Mark Y. Fang, Balaji Sundararaman, Steven M. Blue, et al. 2016a. “Robust Transcriptome-Wide Discovery of RNA-Binding Protein Binding Sites with Enhanced CLIP (eCLIP).” *Nature Methods* 13 (6): 508–14.
———. 2016b. “Robust Transcriptome-Wide Discovery of RNA-Binding Protein Binding Sites with Enhanced CLIP (eCLIP).” *Nature Methods* 13 (6): 508–14.
Van Nostrand, Eric L., Gabriel A. Pratt, Brian A. Yee, Emily C. Wheeler, Steven M. Blue, Jasmine Mueller, Samuel S. Park, et al. 2020a. “Principles of RNA Processing from Analysis of Enhanced CLIP Maps for 150 RNA Binding Proteins.” *Genome Biology* 21 (1): 90.
———. 2020b. “Principles of RNA Processing from Analysis of Enhanced CLIP Maps for 150 RNA Binding Proteins.” *Genome Biology* 21 (1): 90.
Varadi, Mihaly, Stephen Anyango, Mandar Deshpande, Sreenath Nair, Cindy Natassia, Galabina Yordanova, David Yuan, et al. 2022. “AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space with High-Accuracy Models.” *Nucleic Acids Research* 50 (D1): D439–44.
Velázquez-Cruz, Alejandro, Blanca Baños-Jaime, Antonio Díaz-Quintana, Miguel A. De la Rosa, and Irene Díaz-Moreno. 2021. “Post-Translational Control of RNA-Binding Proteins and Disease-Related Dysregulation.” *Frontiers in Molecular Biosciences* 8 (April):658852.
Wang, Jingchao, Zhen Wang, Hiroyuki Inuzuka, Wenyi Wei, and Jing Liu. 2023. “PRMT1 Methylates METTL14 to Modulate Its Oncogenic Function.” *Neoplasia* 42 (August):100912.
Wang, Xiao, Zhike Lu, Adrian Gomez, Gary C. Hon, Yanan Yue, Dali Han, Ye Fu, et al. 2014. “N6-Methyladenosine-Dependent Regulation of Messenger RNA Stability.” *Nature* 505 (7481): 117–20.
Wang, Zhaowei, Qiguo Dai, Jinmiao Song, Xiaodong Duan, Hongpeng Yang, and Ziqi Yang. 2022. “Predicting RBP Binding Sites of RNA with High-Order Encoding Features and CNN-BLSTM Hybrid Model.” *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19 (4): 2409–19.
Wang, Zhonghua, Akash Bhattacharya, and Dmitri N. Ivanov. 2015. “Identification of Small-Molecule Inhibitors of the HuR/RNA Interaction Using a Fluorescence Polarization Screening Assay Followed by NMR Validation.” *PloS One* 10 (9): e0138780.
Wassmer, Elsa, Gergely Koppány, Malte Hermes, Sven Diederichs, and Maïwen Caudron-Herger. 2024. “Refining the Pool of RNA-Binding Domains Advances the Classification and Prediction of RNA-Binding Proteins.” *Nucleic Acids Research* 52 (13): 7504–22.
Wigington, Callie P., Jeenah Jung, Emily A. Rye, Sara L. Belauret, Akahne M. Philpot, Yue Feng, Philip J. Santangelo, and Anita H. Corbett. 2015. “Post-Transcriptional Regulation of Programmed Cell Death 4 (PDCD4) mRNA by the RNA-Binding Proteins Human Antigen R (HuR) and T-Cell Intracellular Antigen 1 (TIA1)*.” *The Journal of Biological Chemistry* 290 (6): 3468–87.
Wolfe, Aaron D., Don B. Arnold, and Xiaojiang S. Chen. 2019. “Comparison of RNA Editing Activity of APOBEC1-A1CF and APOBEC1-RBM47 Complexes Reconstituted in HEK293T Cells.” *Journal of Molecular Biology* 431 (7): 1506–17.
Xu, Junfen, Yifeng Fang, Kelie Chen, Sen Li, Sangsang Tang, Yan Ren, Yixuan Cen, et al. 2022. “Single-Cell RNA Sequencing Reveals the Tissue Architecture in Human High-Grade Serous Ovarian Cancer.” *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* 28 (16): 3590–3602.
Yersal, Ozlem, and Sabri Barutca. 2014. “Biological Subtypes of Breast Cancer: Prognostic

and Therapeutic Implications." *World Journal of Clinical Oncology* 5 (3): 412–24.
Younesian, Samareh, Mohammad Hossein Mohammadi, Ommolbanin Younesian, Majid Momeny, Seyed H. Ghaffari, and Davood Bashash. 2024. "DNA Methylation in Human Diseases." *Heliyon* 10 (11): e32366.
Yu, Yi-Tao, and U. Thomas Meier. 2014. "RNA-Guided Isomerization of Uridine to Pseudouridine--Pseudouridylation." *RNA Biology* 11 (12): 1483–94.
Zaccara, Sara, Ryan J. Ries, and Samie R. Jaffrey. 2019. "Reading, Writing and Erasing mRNA Methylation." *Nature Reviews. Molecular Cell Biology* 20 (10): 608–24.
Zarnegar, Brian J., Ryan A. Flynn, Ying Shen, Brian T. Do, Howard Y. Chang, and Paul A. Khavari. 2016. "irCLIP Platform for Efficient Characterization of Protein-RNA Interactions." *Nature Methods* 13 (6): 489–92.
Zeleznik, Oana A., A. Heather Eliassen, Peter Kraft, Elizabeth M. Poole, Bernard A. Rosner, Sarah Jeanfavre, Amy A. Deik, et al. 2020. "A Prospective Analysis of Circulating Plasma Metabolites Associated with Ovarian Cancer Risk." *Cancer Research* 80 (6): 1357–67.
Zhu, J., A. Mayeda, and A. R. Krainer. 2001. "Exon Identity Established through Differential Antagonism between Exonic Splicing Silencer-Bound hnRNP A1 and Enhancer-Bound SR Proteins." *Molecular Cell* 8 (6): 1351–61.
Zou, Zhongyu, and Chuan He. 2024. "The YTHDF Proteins Display Distinct Cellular Functions on m6A-Modified RNA." *Trends in Biochemical Sciences* 49 (7): 611–21.
Zou, Zhongyu, Caraline Sepich-Poore, Xiaoming Zhou, Jiangbo Wei, and Chuan He. 2023. "The Mechanism Underlying Redundant Functions of the YTHDF Proteins." *Genome Biology* 24 (1): 17.