THE UNIVERSITY OF CHICAGO

RNA MODIFICATIONS: IMPACT ON PROTEIN BINDING, CO-TRANSCRIPTIONAL REGULATION, AND SEQUENCING SIGNATURES

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

INTERDISCIPLINARY SCIENTIST TRAINING PROGRAM:
BIOCHEMISTRY AND MOLECULAR BIOPHYSICS

BY
KATHERINE ISMEI ZHOU

CHICAGO, ILLINOIS
JUNE 2018

Table of Contents

List of Figures

List of Tables

Acknowledgments

I am very grateful to all past and present members of the Pan lab, and to my advisor, Dr. Tao Pan. I would also like to thank the members of my committee: Dr. D. Allan Drummond (chair), Dr. Joseph A. Piccirilli, and Dr. Tobin R. Sosnick.

Many people contributed to this work. I would like to acknowledge Dr. Nian Liu for his initial work on hnRNPC and hnRNPG; Dr. Marc Parisien, Dr. Wesley C. Clark, David W. Pan, and Dr. Matthew J. Eckwahl for bioinformatic analysis; Dr. Qing Dai for chemical synthesis of RNA oligos; Żaneta Matuszek for her work on hnRNPG; Jessica Pan for assistance with hnRNPG cloning; and Dr. Joseph R. Sachleben for technical assistance with nuclear magnetic resonance (NMR). I would like to thank the members of Dr. Demet Araç-Özkan's lab, especially Dr. Gabriel Salzman, for sharing resources and for teaching me to purify proteins from insect cells. I would also like to thank Dr. Jingyi Fei for discussion on RNA polymerase II, and for her advice on microscopy experiments. Finally, many thanks to the University of Chicago NMR Facility, BioPhysics Core Facility, Integrated Light Microscopy Core Facility, Advanced Electron Microscopy Facility, and Genomics Facility for assistance with NMR, biophysics, microscopy, electron microscopy, and sequencing experiments.

This work was supported by the National Institutes of Health Medical Scientist Training Program grant T32GM007281 and F30GM120917. I would also like to thank the generous support of the University of Chicago Biological Sciences Division, the Medical Scientist Training Program, and the Frank Family Endowment.

This thesis includes text and figures derived from articles published in the *Journal of Molecular Biology* by Elsevier [1], in *Molecular Cell* by Cell Press [2], in *Nucleic Acids Research* by Oxford University Press [3], in *Nature Chemical Biology* by SpringerNature [4], and in *RNA Biology* by Taylor & Francis [5].

Abstract

Gene expression is controlled by a complex and interconnected regulatory system. In this thesis, I explore how RNA modifications impact protein binding, co-transcriptional regulation, and sequencing signatures. The cellular functions of the abundant messenger RNA (mRNA) modification N^6-methyladenosine (m^6A) are mediated by m^6A reader proteins. Some m^6A reader proteins selectively bind m^6A-modified RNAs by recognizing motifs that become accessible upon an m^6A-induced change in RNA structure. In Chapter 2, I use biophysical methods to examine how m^6A modification alters the structure of an RNA hairpin to enhance binding of the protein heterogeneous nuclear ribonucleoprotein C (hnRNPC). In Chapter 3, I describe the discovery of another m^6A reader protein, hnRNPG, which binds to a motif that becomes exposed upon m^6A modification of an RNA hairpin. Unlike hnRNPC, hnRNPG binds to a motif that includes the m^6A site. Moreover, hnRNPG uses Arg-Gly-Gly (RGG) repeats in a low-complexity region to selectively bind to m^6A-modified RNA. In Chapter 4, I explore the cellular functions of hnRNPG binding to m^6A-modified RNAs. The hnRNPG protein binds to the phosphorylated C-terminal domain of RNA polymerase II and to nascent m^6A-modified RNA for co-transcriptional, m^6A-dependent gene regulation. In Chapter 5, I build on existing sequencing methods for the detection of another RNA modification, pseudouridine (Ψ). This work spans a variety of perspectives on the impact of RNA modifications, from a biophysical study of the effect of m^6A on RNA structure and protein binding, to an investigation of the cellular functions of m^6A in gene regulation, to method development for the detection of pseudouridines in high-throughput sequencing data. Together, these diverse perspectives demonstrate the widespread impact of RNA modifications.

Chapter 1

Introduction

The regulation of gene activity is crucial for every biological process. Gene activity is regulated on multiple levels: transcriptional, post-transcriptional, translational, and post-translational. While seemingly distinct, these regulatory levels are actually interconnected. Events that occur during transcription can commit an RNA transcript to certain post-transcriptional fates, with far-reaching effects on the cellular localization, translation efficiency, or half-life of the transcribed RNA [6]. Conversely, RNA and protein products feed back onto early steps in their biogenesis, and these feedback mechanisms are central to such fundamental and diverse processes as cell differentiation [7], circadian rhythms [8], and adaptation to stress [9].

In this work, I have studied the modification and splicing of RNAs, as well as RNA–protein interactions. In many cases, these 'post-transcriptional' processes actually occur concurrently with transcription, *i.e.* co-transcriptionally. I have explored how these co-transcriptional events interact with one another and with the process of transcription, both on a molecular level and in the cell.

1.1 Co-Transcriptional Coupling by the Carboxy-Terminal Domain of RNA Polymerase II

In eukaryotes, the coupling of transcription to co-transcriptional processes is facilitated by the carboxy-terminal domain (CTD) of RNA polymerase II (RNAPII) [10–12]. The CTD is a conserved, intrinsically disordered region of the catalytic subunit of RNAPII, located near the RNA exit channel [13]. In mammals, the CTD consists of 52 imperfect tandem repeats of heptapeptides with the consensus sequence Tyr1-Ser2-Pro3-Thr4-Ser5-Pro6-Ser7 [13]. These heptapeptide repeats undergo various post-translational modifications, including phosphorylation [14]. Cyclin-dependent kinase 7 (CDK7), a component of the general transcription

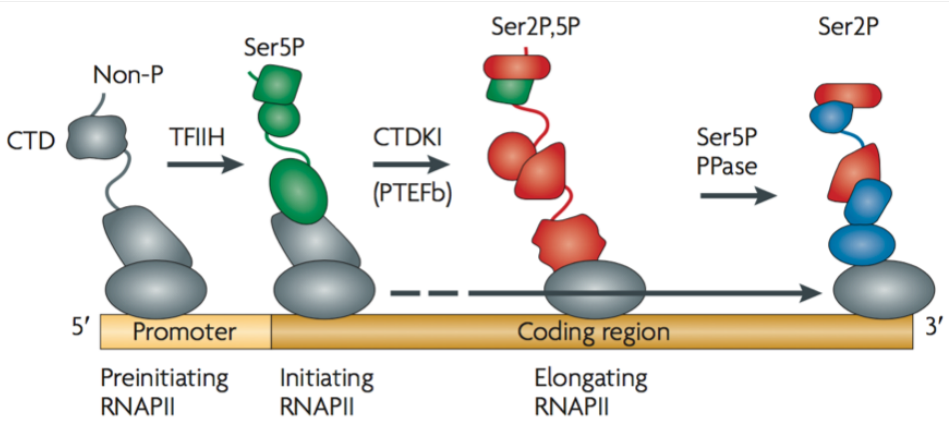


Figure 1.1: The phosphorylation states of the RNAPII CTD during transcription. The nonphosphorylated CTD (Non-P) is associated with pre-initiating RNAPII. TFIIH phosphorylates Ser5 of the CTD (Ser5P) during the transition to initiating RNAPII. CTD kinase I (CTDKI) is the yeast ortholog of CDK9, which is a component of P-TEFb. CTDKI (CDK9 in mammals) phosphorylates Ser2 of the CTD (Ser2P) and generates the actively elongating form of RNAPII. Near the 3′ end of genes, phosphorylation at Ser5 is removed by a protein phosphatase (PPase), and CTD phosphorylation at Ser2 dominates. Gray: nonphosphorylated repeats; green: Ser5P repeats; red: Ser2P,5P repeats (doubly phosphorylated); blue: Ser2P repeats. Proteins bound to a type of repeat are indicated in the same color as the repeat. Source: [19], modified from [20].

factor TFIIH, phosphorylates Ser5 during promoter escape [15], while CDK9, a component of positive transcription elongation factor B (P-TEFb), phosphorylates Ser2 upon promoter-proximal pause release [16]. The different phosphorylation states of the CTD correlate with the different stages of transcription (Figure 1.1): nonphosphorylated CTD predominates in the pre-initiation complex [17], Ser5 phosphorylation (S5P) is associated with early elongation at the 5′ end of the gene, and Ser2 phosphorylation (S2P) is associated with productive elongation and termination, increasing toward the 3′ end of the gene [18].

The CTD of RNAPII is required for the transcription of endogenous genes. Mammalian cells expressing only a CTD-less form of RNAPII can transcribe genes on transiently transfected vectors, albeit with decreased efficiency, but fail to transcribe detectable levels of endogenous genes [12, 21]. This distinction is likely due to differences in chromatin environment and can be explained by the role of the CTD in recruiting factors that modify and remodel chromatin [12]. In addition to this critical role in transcription from native chro-

matin, the CTD is essential for the efficient capping and processing of nascent transcripts [12]. The various phosphorylation states of the CTD coordinate these co-transcriptional processes with the stages of transcription: S5P-CTD recruits $5'$ capping factors at the $5'$ end of the gene [18]; S2P-CTD recruits constitutive splicing factors in the gene body [22–24], as well as $3'$ cleavage and polyadenylation factors at the $3'$ end of the gene [25, 26]. Supporting the importance of the dynamic phosphorylation of the CTD, mutating all Ser5 or Ser2 residues of the CTD to alanines is lethal in budding yeast [27]. In fission yeast, the lethal phenotype caused by mutating all the Ser5 residues of the CTD to alanines can be rescued by tethering the capping complex to RNAPII, demonstrating that the essential role of the Ser5 residues in fission yeast is the recruitment of the capping complex [28]. Mutating the Ser2 residues of the CTD, while not lethal, decreases the recruitment of spliceosome components and $3'$-end cleavage factors to transcription sites in mammalian cells [24].

In addition to its dynamic phosphorylation state, the length and flexibility of the CTD likely play important roles in its function as a landing platform for factors that interact with RNAPII, nascent RNA, and chromatin. The length of the CTD is variable across species, and only $\sim$50% of the repeats in the natural CTD are required for viability in various species from budding yeast to mice [11]. However, CTD length is highly conserved among individuals of the same species [13]. Moreover, while a short CTD is sufficient for RNA processing, CTD length correlates with $3'$-end cleavage activity, so a longer CTD could provide an advantage by increasing RNA processing efficiency [13, 29]. The unbound CTD is intrinsically disordered and structurally compact but heterogeneous [30]. In structures with CTD-binding proteins, CTD repeats often form a β-turn but also adopt a variety of different loop conformations depending on the interacting protein [14]. CTD-binding domains are structurally diverse and use different approaches to recognize CTD repeats with the same modifications [14]. Some proteins that selectively bind to the phosphorylated CTD do not form direct contacts with the phosphate, and their binding might instead be mediated by changes in CTD accessibility and stiffness that occur upon phosphorylation [30].

1.2 Transcription-Coupled Splicing

Both constitutive and alternative splicing can occur concurrently with transcription. RNAPII transcription occurs at a rate of 1.8–4.0 kilobases per minute (30–67 nucleotides per second), and intron half-lives have been estimated at 0.4–7 minutes, consistent with the possibility of co-transcriptional splicing [31]. A single-molecule RNA sequencing study found that, in budding yeast, splicing occurs much more quickly than these estimates suggest, with splicing onset and completion occurring while RNAPII transcribes 26–129 nucleotides (nt) downstream of the $3'$ splice site [32]. Since the RNAPII exit channel covers 15 nt of nascent RNA and the spliceosome binds an estimated 9 nt downstream of the $3'$ splice site, this result suggests that splicing occurs immediately after introns emerge from the RNAPII exit channel [32]. Further supporting co-transcriptional splicing, sequencing of chromatin-associated RNAs in human, mouse, and fly samples has shown that exons are enriched over neighboring intron sequences [33–36]. Moreover, native elongating transcript sequencing (NET-seq) in budding yeast and human cells has revealed RNAPII-associated splicing intermediates, which are indicative of co-transcriptional splicing [37–39]. Estimated proportions of co-transcriptional and post-transcriptional splicing vary widely, likely due to variation among different species or cell types, as well as differing experimental or computational approaches [33–36]. Nonetheless, most sequencing studies agree that splicing is mostly completed while nascent RNAs are associated with chromatin [33]. Moreover, even if splicing is completed post-transcriptionally, spliceosome assembly and commitment to splicing might occur during transcription [12].

While co-transcriptional timing does not necessarily imply that splicing is functionally coupled with transcription, studies show that the RNAPII CTD enhances splicing efficiency both *in vitro* and *in vivo* [12]. Spliceosome components accumulate at the transcription sites of intron-containing RNAs, with over 80% of active spliceosomes bound to chromatin in cultured human cells [31]. S2P-CTD recruits the constitutive splicing factors U2 auxiliary factor 65 kDa subunit (U2AF65) and polypyrimidine tract binding protein associated splicing

factor (PSF) [22–24], and S5P-CTD might function to recruit the spliceosome component U1 small nuclear ribonucleoprotein (snRNP) to 5′ splice sites [10, 40]. Moreover, in budding yeast and humans, S5P-containing RNAPII accumulates at the splice sites of actively spliced exons, possibly as a result of RNAPII pausing [38, 40]. Thus, multiple phosphorylated forms of the CTD, and possibly RNAPII pausing, participate in the recruitment of constitutive splicing machinery to sites of active transcription.

Over 90% of human protein-coding genes are alternatively spliced [41], and mutations that affect alternative splicing contribute to up to 50% of human genetic diseases [42]. Alternative splicing is closely regulated by a complex 'splicing code' of *cis*- and *trans*-acting factors [43]. The *cis*-acting factors include the 5′ and 3′ splice sites, the branch point, and the splicing regulatory elements, which are intronic and exonic splicing enhancers and silencers. Splicing regulatory elements are usually binding sites for RNA-binding proteins that function as *trans*-acting factors. These alternative splicing factors can promote exon inclusion or exclusion by directly influencing the local activity of the spliceosome, for instance by recruiting components of the spliceosome [44] or by inhibiting spliceosome binding or dynamics [45–47]. Other mechanisms for the local regulation of alternative splicing include modulating RNA structure [48, 49] or influencing the binding of other alternative splicing factors [43]. The regulation of alternative splicing is never a truly local and isolated event, however. The splicing of any particular exon is dependent on the splicing of its neighboring exons, since nearby splice sites compete with one another for recognition by the splicing machinery [43]. Moreover, changes in cellular levels or activity of spliceosome components can influence certain alternative splicing events [50]. At the same time, alternative splicing regulation depends strongly on local features and can be highly variable from one regulated exon to another. For example, two well-known classes of splicing factors, the serine/arginine-rich (SR) proteins and the heterogeneous nuclear ribonucleoproteins (hnRNPs), act as positive and negative splicing factors, respectively, in several notable examples [43, 44]. However, high-throughput studies have shown that SR proteins and hnRNPs actually promote similar

numbers of exon inclusion and exclusion events [51–53]. Thus, local features, such as the position of protein binding sites relative to an alternatively spliced exon, help determine whether an exon is positively or negatively regulated by a given splicing factor [43].

In addition to its role in constitutive RNA processing, the CTD has been implicated in the regulation of alternative splicing [12, 31, 54]. The coupling of transcription with alternative splicing can occur through two mechanisms: kinetic coupling and spatial coupling [12, 31, 54]. In kinetic coupling, transcription elongation rate alters the balance in the competition between splice sites or regulatory sequences by changing the rate at which these sequences emerge from the polymerase. In spatial coupling, the CTD recruits alternative splicing factors to the transcription machinery, thereby increasing their local concentration and activity. Kinetic coupling and spatial coupling can influence one another, since elongation rate can affect splicing factor recruitment [55], and conversely alternative splicing factors can regulate elongation rate and CTD phosphorylation [56]. Although the CTD is often involved in factor recruitment, few alternative splicing factors have been shown to directly interact with the CTD. For example, the CTD is required for the recruitment of serine and arginine rich splicing factor 3 (SRSF3) and for its inhibitory effect on a cassette exon in a minigene, but it is unclear whether this recruitment occurs through a direct or indirect interaction [57]. Moreover, since the CTD is present during the transcription of any gene by RNAPII, recruitment by the CTD does not fully explain how an alternative splicing factor targets specific splicing events. Local features such as nucleotide sequence or chromatin state might act together with the CTD to recruit or regulate the activity of alternative splicing factors.

1.3 RNA Modifications

RNA modifications are important modulators of the structure and function of cellular RNAs. Over a hundred different types of modifications have been found in cellular RNAs across the three domains of life [58]. While the numerous modifications found in transfer RNA (tRNA)

and ribosomal RNA (rRNA) have been extensively studied, much less is known about the function of the comparatively sparse modifications found in messenger RNA (mRNA) and long noncoding RNA (lncRNA). Five types of internal base modifications have been identified in eukaryotic mRNAs: N^6-methyladenosine (m^6A) [59–61], N^1-methyladenosine (m^1A) [62], 5-methylcytosine (m^5C) [63, 64], 5-hydroxymethylcytosine (hm^5C) [65], and pseudouridine (Ψ) [66–68]. In this thesis, I will present work on m^6A and pseudouridine.

1.3.1 N^6-methyladenosine (m^6A)

N^6-methyladenosine (m^6A) is the most abundant internal modification in eukaryotic mRNAs. Although m^6A has been known to occur in mRNAs since the 1970's [59], the advent of high-throughput m^6A mapping technologies [60, 61], along with the discovery that m^6A modifications are reversible [69], led to an acceleration in the pace of research in this field [58, 70]. Transcriptome-wide mapping has revealed more than 12 000 m^6A sites in over 7 000 genes in the human transcriptome [60, 61]. These m^6A sites occur in a subset of DRm^6ACH motifs (D = A/G/U, R = A/G, H = A/C/U), with a high density of m^6A sites occurring near stop codons and in long internal exons [60, 61, 71]. Consistent with the reversible and dynamic nature of m^6A modifications, m^6A sites span a wide range of modification fractions [72], and cellular conditions such as heat shock can lead to changes in m^6A modification levels and patterns [60, 61]. The protein machinery supporting the dynamic m^6A modification of cellular RNAs consists of m^6A writers, which deposit m^6A, and erasers, which remove m^6A. The influence of m^6A on the fate of methylated RNAs is mediated by m^6A readers. Through the actions of m^6A writers, erasers, and readers, dynamic m^6A modifications add another dimension to the regulation of the life cycle of cellular RNAs (Figure 1.2).

m^6A writers

A single m^6A methyltransferase complex is responsible for making nearly all the m^6A modifications in mRNA. The core of the complex is a heterodimer of two proteins, methyltransferase-

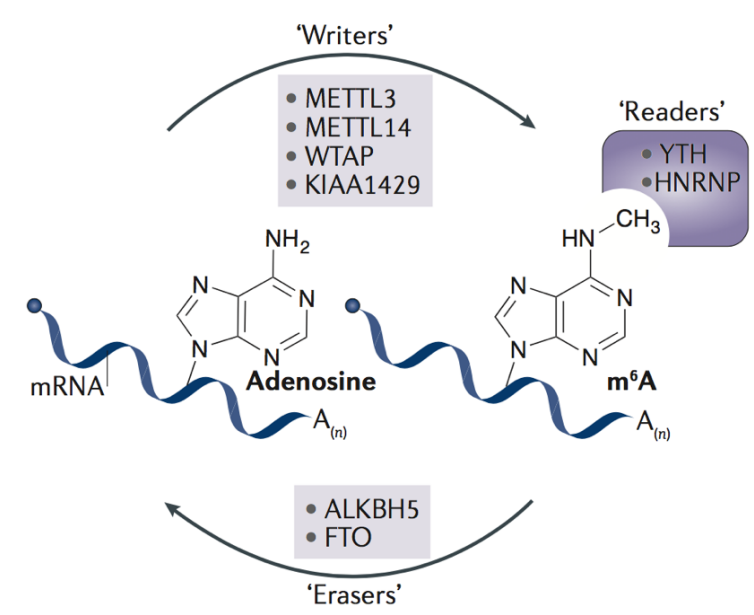


Figure 1.2: The writer, eraser, and reader proteins of m^6A. Source: [73].

like 3 (METTL3) and METTL14. Each of the two proteins has a methyltransferase domain and can methylate RNA *in vitro*, yet the METTL3–METTL14 complex has much greater m^6A methyltransferase activity than either subunit alone [74]. Since both METTL3 and METTL14 have methyltransferase-homologous domains, either subunit could potentially catalyze m^6A methylation in the METTL3–METTL14 complex. However, crystal structures only show densities of the cofactor *S*-adenosylmethionine (SAM) and cofactor product *S*-adenosylhomocysteine (SAH) in the catalytic cavity of METTL3, indicating that METTL3 is the sole catalytic subunit. In contrast, the catalytic cavity of METTL14 adopts a closed conformation that cannot accommodate SAM [75–77]. In addition, mutations in the catalytic site of METTL3 completely abolish the methyltransferase activity of the complex, whereas mutating the putative catalytic site derived from the protein sequence of METTL14 have no effect [75–77], further suggesting that METTL3 is the site of catalysis.

Although METTL3 is the catalytic subunit, its activity is strongly dependent on METTL14. The METTL3–METTL14 complex has much higher activity than the isolated METTL3 protein *in vitro* [74], and knockdown of either METTL3 or METTL14 decreases cellular m^6A

levels to a similar degree [78]. The observation that METTL3 and METTL14 depend on each other for protein stability *in vivo* [78] suggests that METTL14 might enhance the activity of METTL3 in part through structural stabilization. Indeed, crystal structures have shown that METTL3 and METTL14 interact over a large surface area, and mutational analysis has demonstrated the importance of this interface for maintaining solubility of the complex [75–77]. Beyond its role in structural stabilization, METTL14 enhances methyltransferase activity by contributing to RNA substrate recognition through a basic patch close to its interface with METTL3 [75–77]. Thus, while METTL3 is the unique catalytic subunit of the m^6A methyltransferase, METTL14 functions in structural stabilization and RNA substrate recognition.

The METTL3–METTL14 complex deposits m^6A in a sequence-specific manner *in vitro* [74], consistent with the distribution of m^6A at DRm^6ACH motifs transcriptome-wide [60, 61, 71]. A point mutation in METTL14 has been found to decrease both the activity and sequence specificity of the m^6A methyltransferase complex [76]. However, additional sequence specificity elements are likely to be present not only in METTL14, but also in METTL3, since the METTL3 protein alone also prefers to methylate GGACU over GGAUU sequences [74]. Indeed, two zinc finger motifs in METTL3 have been found to preferentially bind a GGACU sequence, although only very weakly, with a binding affinity of several hundred micromolar under physiological conditions [79]. In the METTL3–METTL14 complex, these zinc finger motifs are required for RNA binding and methyltransferase activity [76]. In addition to sequence specificity, the m^6A methyltransferase complex might have intrinsic preferences for RNA substrates with particular secondary structures. Whether the m^6A methyltransferase demonstrates any such preferences *in vitro* or *in vivo* remains unclear.

Additional protein factors that associate with the core METTL3–METTL14 complex *in vivo* have been proposed to recruit the m^6A methyltransferase to its target transcripts. Wilms tumor 1 associated protein (WTAP) associates with METTL3 and METTL14 and is required for both RNA binding and m^6A methyltransferase activity *in vivo* [80, 81]. More-

over, WTAP is necessary for the localization of METTL3 and METTL14 to nuclear speckles [80]. The human protein virilizer homolog, vir-like m^6A methyltransferase associated (VIRMA, also called KIAA1429), is also required for m^6A methyltransferase activity *in vivo* [81] and mediates the preferential deposition of m^6A near stop codons and in the $3'$ UTR [82]. The importance of both WTAP and VIRMA for m^6A deposition is further highlighted by the fact that knockdown of either gene leads to a greater decrease in m^6A levels than knockdown of either METTL3 or METTL14 [81]. In addition to these proteins that interact directly with METTL3 and METTL14, RNA-binding motif protein 15 (RBM15) and RBM15B bind the m^6A methyltransferase complex indirectly through an interaction with WTAP [83]. RBM15 and RBM15B recruit m^6A methyltransferase to specific sites in the lncRNA X-inactive specific transcript (*XIST*) for the targeted deposition of m^6A modifications [83]. It has also been proposed that microRNAs (miRNAs) mediate the targeting of m^6A methyltransferase to specific sites for m^6A deposition, since disrupting miRNA biogenesis or mutating miRNA sequences can affect both METTL3 binding and m^6A levels [84]. However, the mechanisms by which miRNAs might influence targeting of the m^6A methyltransferase remain unclear.

Similar to $5'$-end capping, splicing, and $3'$-end processing, internal modification of RNA can also occur co-transcriptionally. The catalytic subunit of the m^6A methyltransferase, METTL3, associates with chromatin and is thought to deposit m^6A on nascent RNAs [85, 86]. Supporting co-transcriptional deposition, m^6A has been detected in nascent and chromatin-associated RNA [87, 88], and mRNAs transcribed from METTL3-bound genes have higher m^6A levels than those transcribed from nonbound genes [85, 86]. The mechanism by which METTL3 is recruited to chromatin, as well as the precise site of METTL3 binding within target genes, is unclear. While a study in human leukemia cells found that METTL3 was recruited to transcription start sites by the CAATT enhancer binding protein zeta (CEBPZ) [85], a study in murine embryonic stem cells instead found that chromatin-bound METTL3 was enriched at the $3'$ end of protein-coding genes [86]. METTL3 has also

been found to interact with RNAPII in a human breast cancer cell line treated with the topoisomerase inhibitor camptothecin [89]. Since camptothecin stalls elongating RNAPII, this result raises the possibility that m^6A deposition not only occurs co-transcriptionally but also can be influenced by transcriptional elongation [89].

Although the METTL3–METTL14 complex is responsible for most of the m^6A modifications in cellular mRNA, the nuclear protein METTL16 has been shown to deposit m^6A in at least one mRNA [90]. METTL16 also methylates a conserved m^6A site in U6 small nuclear RNA (snRNA) both in fission yeast and in human cells [90]. In contrast to the DRm^6ACH motif preferred by METTL3–METTL14, both m^6A modifications installed by METTL16 occur in a UACm^6AGAGAA context [90]. While over 2 000 other m^6A sites, mostly in introns or near intron–exon boundaries, were found to be METTL16-dependent, it is unclear whether any of these other sites are direct targets of METTL16 [90].

The METTL3–METTL14 m^6A methyltransferase complex is essential for life in mice and is required for normal development in fruit flies [91–93]. Studies in mouse and human embryonic stem cells have demonstrated that METTL3 is required for transcript turnover and stem cell differentiation [78, 91, 94]. Moreover, METTL14 functions in neural stem cell self-renewal and embryonic cortical neurogenesis [95, 96], and METTL3 is essential for the specification of hematopoietic stem/progenitor cells in zebrafish [97]. In addition to these functions in development, METTL3 controls the rate of the circadian clock [98], and both METTL3 and METTL14 have been implicated in multiple types of cancer [99]. While METTL3 and METTL14 suppress tumorigenesis and metastasis in some cancer models, they promote proliferation and invasion in others [99]. Notably, METTL3 promotes the growth and invasion of lung adenocarcinoma cells through mechanisms independent of its m^6A methyltransferase activity [100]. In contrast, the catalytic activity of METTL3 is required for the proliferation of acute myeloid leukemia (AML) blasts [85, 101, 102]. Together, these studies have revealed critical functions for m^6A in development and human disease.

m^6A erasers

The discovery that m^6A modifications are reversible [69] provided part of the impetus for m^6A research. Two known demethylases, fat mass and obesity associated protein (FTO) and AlkB family member 5 (ALKBH5), mediate the reversal of m^6A modifications [69, 103]. Both proteins belong to the AlkB family of Fe(II)- and α-ketoglutarate-dependent dioxygenases. Through the oxidative demethylation of m^6A, FTO generates the long-lived intermediates N^6-hydroxymethyladenosine and N^6-formyladenosine, which are present in cellular RNA and could potentially have additional regulatory functions [104]. Unlike the m^6A methyltransferase complex, FTO and ALKBH5 do not display sequence specificity, although both demethylases bind and demethylate m^6A sites located in single-stranded RNA regions more efficiently than an m^6A site in an RNA duplex [105]. Thus, it is unclear how FTO and ALKBH5 are targeted to specific m^6A sites, or how their activity is regulated. One possibility is that, like the m^6A methyltransferase complex, FTO and ALKBH5 are targeted and regulated by additional protein factors. In addition to demethylating m^6A, FTO removes the N^6-methyl of N^6,2′-*O*-dimethyladenosine (m^6A_m), which occurs as the first nucleotide following the 5′ cap structure in up to 30% of mRNAs [106]. In fact, *in vitro* demethylation by FTO was shown to be 100-fold more efficient for m^6A_m in the context of the 5′ cap than for internal m^6A [106]. In contrast, ALKBH5 does not display any activity toward m^6A_m [106].

Multiple genome-wide association studies (GWAS) found that *FTO* gene variants are associated with obesity and type 2 diabetes [107–109]. However, these variants were later shown to influence body mass by enhancing the expression of the Iroquois homeobox gene *IRX3*, whereas they have no effect on *FTO* expression [110]. Nonetheless, a direct link between FTO and body mass regulation has been demonstrated in mice, where FTO expression positively correlates with body and fat mass [111, 112]. FTO and m^6A have also been implicated in adipogenesis [113] and memory formation [114]. Moreover, FTO influences mRNA splicing and 3′-end processing through its effect on m^6A levels [115]. While

both FTO and ALKBH5 are nuclear proteins that partially co-localize with nuclear speckles [69, 103], in neurons FTO has been found to be enriched in axons, where it regulates the m^6A modification and local translation of mRNAs [116]. The second m^6A demethylase, ALKBH5, functions in pathways that are largely distinct from those of FTO. Loss of ALKBH5 demethylase activity in cultured human cells leads to accelerated mRNA export [103]. In mice, ALKBH5 deficiency causes aberrant mRNA splicing, $3'$ UTR length, and turnover in spermatocytes and round spermatids [117], which leads to defects in spermatogenesis [103]. Finally, both FTO and ALKBH5 have been implicated in multiple types of cancer [99].

m^6A readers

The reversible m^6A modification of an mRNA molecule can impact every step in its life cycle, including mRNA splicing [92, 93, 118, 119], export [120], translation [121, 122], and decay [123–126]. m^6A modifications alter the fate of modified RNAs by influencing their interactions with RNA-binding proteins that selectively bind to m^6A-modified RNAs, called m^6A reader proteins.

The m^6A reader proteins identified to date fall into three classes, sorted based on their mechanisms for selective binding to m^6A-modified RNAs. Class I includes the m^6A reader proteins that contain a YT521-B homology (YTH) domain [127]. In these m^6A readers, the YTH domain directly and selectively binds the m^6A base in a hydrophobic aromatic cage [128–130] (Figure 1.3(a)). Class II includes two m^6A reader proteins so far, both of which are hnRNPs. These m^6A readers selectively bind m^6A-modified RNAs through an m^6A-switch mechanism, in which m^6A decreases the stability of Watson–Crick base pairing and thereby increases the accessibility of a single-stranded RNA binding motif that is recognized by the m^6A reader (Figure 1.3(b)). Class III includes the insulin-like growth factor-2 mRNA-binding proteins 1, 2, and 3 (IGF2BP1–3). In these m^6A readers, common RNA binding domains and their flanking regions act together to selectively bind m^6A-modified transcripts

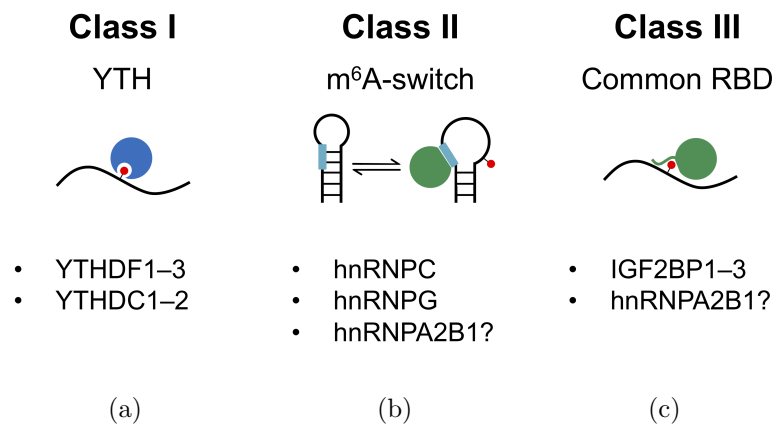


Figure 1.3: Three classes of m^6A reader proteins. (a) Class I m^6A reader proteins use a YTH domain (blue) to bind the m^6A base (red) in an aromatic cage. This class includes the five human proteins that contain a YTH domain. (b) Class II m^6A reader proteins use an m^6A-switch mechanism to bind m^6A-modified transcripts: m^6A modification of RNA destabilizes Watson-Crick base-pairing and increases the accessibility of a single-stranded RNA motif (cyan), which is recognized by the m^6A reader protein (green). This class includes hnRNPC, hnRNPG, and possibly hnRNPA2B1. (c) Class III m^6A reader proteins use a common RNA binding domain (RBD) and its flanking regions (green) to recognize m^6A-modified transcripts. This class includes the IGF2BPs, which use KH domains and their flanking regions to selectively bind m^6A-modified RNAs, and possibly hnRNPA2B1, in which the RRMs and their flanking regions might contribute to m^6A selectivity. Source: [4].

(Figure 1.3(c)). The m^6A reader protein hnRNPA2B1, which also lacks a YTH domain, could fall into either Class II or III. One of the RNA recognition motifs (RRMs) of hnRNPA2B1 might partly account for its m^6A selectivity, possibly by acting in combination with flanking regions [131, 132]. Since RRMs are common RNA binding domains [133], this mechanism would define hnRNPA2B1 as another Class III m^6A reader protein.

Class I: The YTH domain family. Members of the YTH domain family were first identified as m^6A reader proteins in an RNA pull-down from the lysates of cultured human cells, in which the RNA bait contained a known m^6A site in the Rous sarcoma virus (RSV) genome [60]. Five human proteins contain YTH domains: YTH domain family 1, 2, and 3 (YTHDF1–3) and YTH domain-containing 1 and 2 (YTHDC1–2) [127]. The YTH domain directly binds to the m^6A base, with a binding dissociation constant (K_d) for m^6A-containing RNAs in the

100–300 nM range, whereas binding to nonmethylated RNAs is typically 5- to 20-fold weaker [127, 128, 130]. The first YTH domain family protein to be functionally characterized as an m^6A reader was YTHDF2 [123]. While the C-terminal YTH domain selectively bound to m^6A, the N-terminal domain was found to shorten the half-life of an mRNA reporter when tethered to its $3'$ UTR [123]. Further studies showed that YTHDF2 promotes the decay of m^6A-modified transcripts by recruiting the deadenylase complex carbon catabolite repressor 4 (CCR4) – negative on TATA (NOT) [134]. Consistent with the association between m^6A and mRNA decay, knocking out m^6A methyltransferase leads to a global decrease in m^6A levels as well as an increase in the steady-state levels of m^6A-containing mRNAs [91].

YTH domain-containing proteins have been proposed to sort m^6A-containing transcripts into a fast track with accelerated nuclear export, translation, and decay [73]. The nuclear reader YTHDC1 promotes the nuclear export of m^6A-containing RNAs [120], while the cytoplasmic reader YTHDF1 increases the translation efficiency of m^6A-containing mRNAs, both by recruiting them to the translation machinery and by increasing the efficiency of translation initiation [121]. Another cytoplasmic reader, YTHDF3, appears to promote both the translation and decay of m^6A-containing mRNAs by acting cooperatively with YTHDF1 and YTHDF2 [135, 136]. The cytoplasmic reader YTHDC2 has $3'$-to-$5'$ RNA helicase activity *in vitro* and has been proposed to promote mRNA decay by recruiting the $5'$-to-$3'$ exoribonuclease 1 (XRN1) to m^6A-containing RNAs [137].

Some YTH domain-containing proteins appear to have multiple functions. In response to heat shock, the normally cytoplasmic YTHDF2 protein re-localizes to the nucleus, where it binds to m^6A sites in the $5'$ untranslated regions (UTRs) of stress-induced RNAs and protects them from demethylation [122]. Since m^6A modifications in the $5'$ UTR can mediate cap-independent translation [138], the protection of these modifications by YTHDF2 allows them to be translated as part of the heat shock stress response. In addition to promoting nuclear export, YTHDC1 has functions in the regulation of transcription and splicing. YTHDC1 binds to m^6A sites in the lncRNA *XIST* and is required for transcriptional silencing of the

X chromosome in cultured human cells [83]. In addition, YTHDC1 promotes the inclusion of hundreds of exons by modulating the recruitment of the serine and arginine rich splicing factors SRSF3 and SRSF10 to m^{6}A-containing transcripts [118]. Interestingly, YT521-B, the only nuclear YTH family protein in *Drosophila melanogaster*, recognizes m^{6}A and promotes the female-specific alternative splicing of the sex determination factor *Sex lethal* [92, 93].

Class II: The m^{6}A-switch mechanism. The nuclear protein hnRNPC was identified as an m^{6}A reader that lacks a YTH domain. Instead, the recognition of m^{6}A by hnRNPC depends on an m^{6}A-induced change in RNA structure [119]. While m^{6}A is capable of Watson–Crick base pairing, thermal denaturation studies with model RNA duplexes have demonstrated that m^{6}A in a duplex is destabilizing by 0.5–1.7 kcal/mol [139, 140]. To allow hydrogen bonding at the Watson–Crick face, the N^6-methyl is rotated such that it is in the *anti* position relative to the N1 across the C6–N6 bond [139]. The destabilization of the duplex by m^{6}A methylation is likely due to the steric clash between N7 and the *anti* N^6-methyl in base-paired m^{6}A [139]. In addition to these *in vitro* studies, transcriptome-wide mapping of RNA structure *in vivo* has shown that m^{6}A sites are less structured than other sites with the same sequence [141]. This structural signature results from m^{6}A-induced changes in RNA structure, and not from the structural specificity of m^{6}A methyltransferases, since the signature disappears along with m^{6}A modification upon knockdown of the m^{6}A methyltransferase component METTL3 [141].

Since hnRNPC binds single-stranded U-tract motifs, methylation of an adenosine in a hairpin-stem can destabilize the duplex to expose an hnRNPC binding site [142, 143]. This 'm^{6}A-switch' mechanism was found to be the basis by which hnRNPC recognizes m^{6}A modification at a site in the human lncRNA metastasis-associated lung adenocarcinoma transcript 1 (*MALAT1*). hnRNPC binds the m^{6}A-containing site with a K_d of 90–100 nM, which corresponds to an ∼8-fold higher affinity than hnRNPC binding to the nonmethylated site [119]. Further examination of hnRNPC-bound RNAs revealed 2 798 high-confidence

m^6A switches in which hnRNPC is thought to use a similar mechanism of indirect m^6A recognition [119]. In addition, hnRNPC binding to these m^6A switches was found to regulate the abundance of thousands of mRNA transcripts and the alternative splicing of hundreds of alternative exons [119]. Given the impact of RNA structure on protein binding, m^6A-switch-like changes in RNA structure likely have a widespread effect on the binding of many proteins to m^6A-modified RNAs. In Chapter 3, I present work showing that the protein hnRNPG also uses an m^6A-switch mechanism to selectively bind m^6A-containing RNAs [3].

The impact of m^6A on RNA structure could conceivably also have functional consequences on methylated RNAs without affecting protein binding. For instance, m^6A could impact miRNA–mRNA interactions, especially since m^6A modifications are commonly found in the 3′ UTR of mRNAs [144]. Since m^6A is mostly deposited co-transcriptionally, m^6A modifications also have the potential to alter co-transcriptional splicing by influencing pre-mRNA–snRNA interactions, and could also affect co-transcriptional RNA folding, which can in turn impact transcription and co-transcriptional processing [145].

Class III: Common RNA binding domains and flanking unstructured regions. One of the well-established functions of m^6A modification is to promote mRNA decay through a mechanism mediated by YTHDF2 [123]. However, overexpression of an m^6A demethylase reduces both the m^6A content and the steady-state levels of hundreds of transcripts, and some of these transcripts have been shown to be more stable in the presence of m^6A modification [146]. This observation led to the discovery that insulin-like growth factor-2 mRNA-binding proteins 1, 2, and 3 (IGF2BP1–3) are m^6A reader proteins that stabilize their m^6A-containing target transcripts [147]. IGF2BP1–3 lack a YTH domain and bind to m^6A-containing RNA independently of RNA structure [147], suggesting that they use a third mechanism to recognize m^6A-containing transcripts. The IGF2BPs are composed of two RRMs and four hnRNPK-homology (KH) domains, and only the third and fourth KH domains (KH3–4) are required for the selective recognition of m^6A-containing RNAs

[147]. Whereas the m^{6}A-selective YTH domains are only present in five human proteins [127], KH domains are among the most common RNA binding domains, found in dozens of human RNA-binding proteins [133]. It is surprising that the common KH domains could be responsible for the selective binding of IGF2BP1–3 to m^{6}A-containing transcripts in an RNA-structure-independent manner. However, canonical RNA-binding domains can use non-canonical modes to bind RNA. For instance, some RRMs do not use the canonical β-sheet binding surface for RNA recognition, but instead rely entirely on the loops between the α-helices and β-strands of the RRM domain [148]. Similarly, the KH domains of IGF2BP1–3 might use unconventional surfaces or loops, or flanking disordered linker regions, to selectively bind m^{6}A-containing RNA. In fact, the KH3–4 di-domains of IGF2BP1–3 are insufficient for selective binding to m^{6}A-containing RNAs on their own, and it has been proposed that regions flanking the KH domains also contribute to m^{6}A selectivity [147]. Thus, while the exact mechanisms of m^{6}A selectivity have yet to be worked out, IGF2BP1–3 establish a third class of m^{6}A reader proteins (Figure 1.3(c)).

In addition to the known Class II and III readers, other m^{6}A reader proteins lacking a YTH domain have been shown to regulate primary-microRNA (pri-miRNA) processing and mRNA translation [131, 138]. These reader proteins might recognize m^{6}A-modified RNAs through m^{6}A-induced changes in RNA structure (Class II), by directly binding the m^{6}A base with common RNA binding domains and their flanking regions (Class III), or using as-yet-unknown mechanisms.

m^{6}A dynamics and functions: future directions

The RNA modification m^{6}A has a wide variety of functions in both coding and noncoding RNAs, which have been described in multiple recent review papers [99, 127, 149]. Here, I will briefly discuss a few open questions and developing areas of the m^{6}A field.

m^6A dynamics. One of the intriguing features of m^6A is that it can be dynamic and reversible. Although one study found that m^6A modifications do not substantially change between chromatin-associated and mature mRNA [88], changes in the distribution and levels of m^6A modifications have been observed in response to stress [122, 150], during development [151], and in disease [99]. However, so far the regulation of m^6A modification levels has primarily been explained by global changes in methyltransferase or demethylase levels or activity, and the mechanisms behind the regulation of m^6A modification at specific sites or even in specific transcripts remain unclear.

Heat shock leads to increased levels of m^6A modification in the $5'$ UTR of stress-induced mRNAs through the protection of these m^6A sites from demethylation. Although the increased protection of these sites is explained by re-localization of YTHDF2 to the nucleus, it is unclear how YTHDF2 specifically protects m^6A sites in the $5'$ UTRs of stress-induced mRNAs, as opposed to other m^6A sites. Another stress-induced change in m^6A modification levels occurs in response to ultraviolet light-induced DNA damage [150]. m^6A modification levels transiently increase at sites of DNA damage, possibly through the recruitment of METTL3 followed by FTO. However, the mechanism behind the recruitment of these factors to sites of DNA damage is unknown. Given that METTL3 interacts with RNAPII upon camptothecin treatment [89], one possibility is that RNAPII stalling at sites of DNA damage plays a role in the recruitment of METTL3. Although transiently methylated transcripts recruit a DNA polymerase to sites of damage, the identities of the RNA transcripts that are methylated in response to DNA damage, as well as the fate of the methylated transcripts, are also unclear. For example, it is possible that all transcripts in the vicinity of sites of DNA damage are methylated, that m^6A is specifically installed on abortive transcripts generated upon RNAPII arrest due to DNA damage, or that the m^6A-modified RNAs have additional functions in the DNA damage response beyond the recruitment of DNA repair machinery.

m^6A in co-transcriptional regulation. Since m^6A is mostly deposited co-transcriptionally [88], m^6A modifications have the potential to regulate co-transcriptional processes. Indeed, m^6A has been implicated in alternative polyadenylation site selection [152] and has been shown to promote the processing of pri-miRNAs [153], which can occur co-transcriptionally and is coupled to transcription by RNAPII [154–156]. Another primarily co-transcriptional process is splicing [31]. Supporting a role for m^6A in alternative splicing, the m^6A methyltransferase complex co-localizes with splicing factors in nuclear speckles, and ∼30% of m^6A methyltransferase binding sites are found in introns [74]. Furthermore, m^6A is enriched in long internal exons [60], in the last exon of RNA transcripts [88], and in exonic regions near splice junctions [87, 113]. In addition, increased levels of m^6A in exonic regions near splice junctions are associated with greater splicing efficiency [87], and m^6A methyltransferase depletion studies demonstrate an impact of m^6A on alternative splicing [119]. The m^6A reader proteins YTHDC1 and hnRNPC have been shown to regulate alternative splicing in an m^6A-dependent manner [118, 119]. In Chapter 4, I will describe how another m^6A reader protein, hnRNPG, interacts with both m^6A-modified nascent RNA and the transcriptional machinery for co-transcriptional, m^6A-dependent alternative splicing regulation.

Position-dependent effects of m^6A. Various patterns in the distribution of m^6A along transcripts have been noted: m^6A is abundant near the stop codon and in the 3′ UTR [60, 61], in long internal exons and in the last exon [60, 152], and in exons near splice junctions [87, 113]. However, the link between the position and function of m^6A modifications remains unclear. For instance, the m^6A reader protein YTHDF2 can bind m^6A modifications in either the coding sequence or 3′ UTR to stimulate mRNA turnover [123, 134]. The functional distinction between YTHDF2-bound m^6A sites in the coding region and the 3′ UTR, if any, is unknown. Furthermore, in addition to YTHDF2, YTHDF1 can bind in the 3′ UTR to stimulate mRNA translation [121], and IGF2BP1–3 can bind in the 3′ UTR to promote mRNA stability [147]. It remains unknown how these different m^6A readers might compete

to bind the same methylated 3′ UTRs. This question is particularly relevant for YTHDF2 and IGF2BP1–3, given that these m^{6}A readers have opposing effects on mRNA stability. While m^{6}A modifications in exonic regions near splice junctions are associated with more efficient splicing [87], the mechanism behind this position-dependent association has yet to be determined. Given that the distribution of m^{6}A modifications along RNA transcripts is not uniform, it is likely that the positions of m^{6}A sites have some functional impact.

Regulation of proteins by m^{6}A While interactions between m^{6}A-modified transcripts and m^{6}A reader proteins have primarily been described to alter the fate of the m^{6}A-modified RNAs, these interactions can also influence the fate of the m^{6}A reader proteins. m^{6}A sites can recruit m^{6}A reader proteins to particular sites of action: m^{6}A-modifed *XIST* recruits YTHDC1, which is required for transcriptional repression of the X chromosome [83]; m^{6}A modifications recruit DNA polymerase κ to sites of DNA damage through an unknown intermediary m^{6}A reader [150]; and in Chapter 4 I will describe how m^{6}A-modified nascent RNA recruits hnRNPG to transcription sites to regulate alternative splicing. In addition to regulating protein localization, m^{6}A might impact m^{6}A reader protein conformation, assembly, or activity. Through these and other mechanisms, just as some RNA-binding proteins might be regulated by RNA instead of regulating RNA [157], some m^{6}A reader proteins might be regulated by m^{6}A-modified RNA instead of, or in addition to, regulating m^{6}A-modified RNA.

1.3.2 Pseudouridine (Ψ)

Pseudouridine (Ψ) is the most abundant modification in cellular RNAs [158, 159]. In contrast to m^{6}A, pseudouridine can be installed by multiple different enzymes. Either RNA-dependent or RNA-independent enzymes can isomerize uridine (U) to generate pseudouridine at specific sites in noncoding and coding RNAs, including rRNA, tRNA, snRNA, and mRNA [158–161]. RNA-dependent pseudouridylation is carried out by box H/ACA small

nucleolar ribonucleoproteins (snoRNPs), in which the snoRNA guides the enzyme to the target sequence. RNA-independent pseudouridylation is carried out by pseudouridine synthase (Pus) enzymes that target specific RNA structures or sequences. Since $2'$-O-methylation of rRNA occurs co-transcriptionally, it has been assumed that pseudouridylation of rRNA also occurs co-transcriptionally [162]. However, it is unclear whether mRNAs are pseudouridylated co-transcriptionally, and in at least some cases they are likely pseudouridylated post-transcriptionally in the cytoplasm [66].

Pseudouridines in noncoding RNAs are conserved and cluster in functionally important regions, such as the peptidyl transferase center in rRNAs and the branch-site recognition sequence in U2 snRNA [160]. By increasing phosphodiester backbone rigidity and base stacking [163–165], pseudouridines impact intramolecular and intermolecular RNA–RNA interactions, including rRNA folding, tRNA binding to the ribosome, and pre-mRNA–snRNA base pairing [166–168]. Although pseudouridines in mRNAs are also conserved [66, 67], their functions are less well understood. The introduction of pseudouridines into stop codons can decrease the efficiency of translation termination [169], and pseudouridine has also been proposed to alter the decoding of sense codons [170]. Moreover, the pseudouridylation of >200 sites in mRNA is induced upon heat shock in budding yeast, and it has been proposed that pseudouridine functions to stabilize these modified mRNAs [66]. An impact of pseudouridine on mRNA stability might depend on pseudouridine reader proteins, or the stabilization of RNA structure by pseudouridine might inhibit the degradation of pseudouridine-containing mRNAs by the exosome complex [171].

Pseudouridylation does not occur to completion at all pseudouridine sites [58, 163, 172]. Most pseudouridine sites in mRNA as well as certain sites in rRNA are only partially modified [58, 163, 172]. In addition, pseudouridylation at specific sites in rRNA, snRNA, and mRNA can be induced in response to changes in growth conditions or stress [66, 68, 173]. For instance, two sites in U6 snRNA are induced upon nutrient deprivation in budding yeast, and these additional pseudouridine sites have been proposed to function in the regulation of

pre-mRNA splicing [173]. Although stress-induced pseudouridylation seems to occur at sites that differ from the consensus target sequences of snoRNA guides or Pus enzymes, the mechanisms by which the specificity of pseudouridylases is relaxed upon stress remain unclear. No pseudouridine de-modification enzyme has been identified, and given the stability of the C–C glycosidic bond, pseudouridine modifications are likely irreversible on a nucleoside level [174]. However, pseudouridines are at least passively reversible on a transcript level through RNA turnover.

No pseudouridine reader proteins have been discovered so far. The effect of pseudouridine on RNA structure could potentially influence the binding of many RNA-binding proteins through a switch-like mechanism. Unlike m^6A, pseudouridine mostly stabilizes RNA structure, so RNA-binding proteins that use this mechanism to selectively bind pseudouridine-containing RNAs would be expected to preferentially bind structured RNAs. Alternatively, pseudouridine reader proteins could potentially distinguish directly between the pseudouridine and uridine bases, for instance by recognizing the extra hydrogen-bond donor at the N1 position of pseudouridine. The human single-stranded RNA-binding protein Pumilio 2 (hPUM2) was found to act as an 'anti-reader' of both pseudouridine and m^6A, in that its interaction with its consensus binding site is reduced upon pseudouridine or m^6A modification [175]. Other such 'anti-readers' have been identified for m^6A and might also exist for pseudouridine [176, 177].

Pseudouridine sites in cellular RNAs can be detected through the selective chemical modification of pseudouridines with *N*-cyclohexyl N'-(2-morpholinoethyl) carbodiimide (CMC) [178]. CMC reacts with pseudouridine, uridine, and guanosine, but only the CMC-modified N3 of pseudouridine is resistant to alkaline hydrolysis (Figure 1.4). Following CMC treatment, pseudouridine sites can be identified by primer extension, since reverse transcription terminates one nucleotide $3'$ to CMC-modified pseudouridine (CMC-Ψ). The combination of this method with high-throughput sequencing revealed hundreds of pseudouridine sites in mRNAs and noncoding RNAs previously thought to lack pseudouridylation [66–68]. The

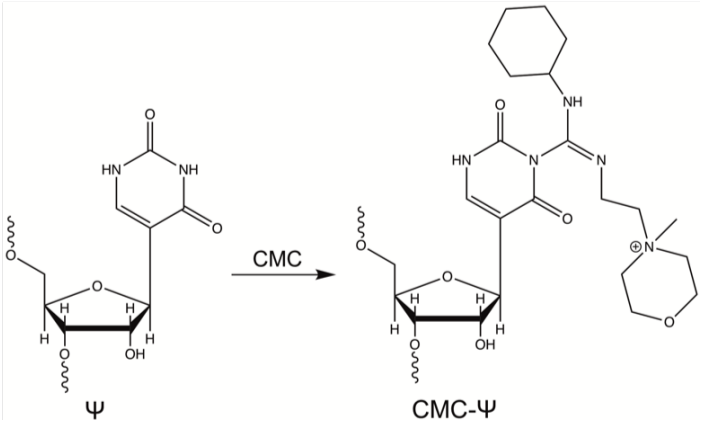


Figure 1.4: Chemical structures of Ψ and CMC-modified Ψ (CMC-Ψ).

sensitivity of sequencing-based pseudouridine detection was further improved by using a chemically synthesized derivative of CMC that allowed the biotinylation and enrichment of pseudouridine-modified transcripts prior to reverse transcription [179]. These studies also identified several novel pseudouridine sites in rRNA, of which three were validated through direct detection of the pseudouridine nucleoside using site-specific cleavage and radioactive-labeling followed by ligation-assisted extraction and thin-layer chromatography (SCARLET) [68, 72, 179]. Together, these studies greatly expanded the scope of known pseudouridine modifications in eukaryotic RNAs.

1.3.3 Transcriptome-wide distribution of RNA modifications

The development of high-throughput sequencing methods to map RNA modifications transcriptome-wide, followed by the application of these methods to diverse systems, has led to important insights on the distribution and function of RNA modifications. RNA modifications have been mapped in various RNA species in all three domains of life: in addition to being found in abundant noncoding RNAs in eukaryotes, eubacteria, and archaebacteria, m^6A modifications have been identified in the mRNAs of both eukaryotes and eubacteria, and pseudouridine has been found in eukaryotic mRNAs and eubacterial Y RNAs [180]. In addition, m^6A modifications have been mapped in a variety of mammalian tissues, during

different stages of cell differentiation and development, in diseases including cancer, and in viruses [149]. Despite these significant advances, the RNA modification field has faced major challenges in the development of mapping methods that are simultaneously sensitive, specific, high-resolution, high-throughput, and quantitative.

Current methods for the high-throughput mapping of RNA modifications are based on the selectivity of either antibodies or chemical treatments [181, 182]. In antibody-based methods, RNA fragments that contain the modification of interest are selectively enriched using an antibody that binds the modified nucleoside with high specificity and affinity. This enrichment is the basis of existing methods for mapping m^6A [60, 61]. However, antibody-based methods face several challenges. First, they depend on the existence of an antibody with high affinity and selectivity for the modified nucleoside, which has been a challenge for other modifications such as pseudouridine. Second, these methods are subject to antibody artifacts and variability. For instance, half of all m^6A sites identified using antibody-based sequencing were found to be independent of m^6A methyltransferase [183]. Third, particularly since antibodies are large proteins, their recognition of RNA modifications can vary depending on structural or sequence context. Fourth, without additional cross-linking steps, antibody-based methods can only achieve a low resolution, identifying RNA modifications that occur within a $\sim$100 nucleotide range [60, 61]. Together, these shortcomings pose limitations to sensitivity, specificity, and resolution, and further prevent antibody-based methods from being used to quantitatively determine RNA modification fraction.

In chemical-based methods, cellular RNA is treated with a chemical that can distinguish between modified and unmodified nucleosides, and nucleosides that have reacted with the chemical can subsequently be identified by their stop or mutation signatures in high-throughput sequencing. This approach is the basis of pseudouridine sequencing methods, which rely on the selective reaction of CMC with pseudouridines to create a bulky adduct that leads to reverse transcription stops [66–68], as well as m^5C sequencing methods, which rely on the selective reaction of bisulfite with unmodified cytosines to create uracils that can

be identified as mutations in high-throughput sequencing [63, 64]. These methods face several challenges. First, the chemical treatment is often damaging to RNA, leading to artifacts as well as a requirement for large input samples. Second, chemical modification and reversal can often be incomplete, leading to a lack of sensitivity and specificity, which has been a major problem for the detection of both pseudouridine and m^5C [181, 184]. Third, reliance on reverse transcription stops, which applies to pseudouridine sequencing methods, leads to severe limitations in specificity and sensitivity, due to the wide range of possible causes of reverse transcription stops as well as the inability to recognize modifications that are close together or close to the $3'$ end of RNA. On the other hand, compared to antibody-based methods, chemical-based methods are less likely to be highly dependent on structural and sequence context, since chemicals are much smaller than antibodies and the RNA can usually be denatured for chemical treatment. Moreover, chemical-based methods can often achieve single-nucleotide resolution and tend to be at least semi-quantitative. Still, so far neither antibody- nor chemical-based methods have been able to accurately measure absolute RNA modification fractions. Moreover, both types of methods are designed to target a specific RNA modification. Thus, they cannot detect multiple different types of modifications at once, and can in fact fail to distinguish between different types of RNA modifications [181].

The combination of multiple approaches has led to improved methods for the high-throughput mapping of RNA modifications [182]. For instance, photo-cross-linking-assisted m^6A sequencing (PA-m^6A-seq) [185] and m^6A individual-nucleotide-resolution cross-linking and immunoprecipitation (miCLIP) [71] have combined antibody-based enrichment of m^6A-modified RNAs with cross-linking to greatly improve the resolution of m^6A site mapping. Conversely, in N_3-CMC-enriched pseudouridine sequencing (CeU-seq), a biotinylated carbodiimide is used to enrich pseudouridine-containing RNAs to enhance the sensitivity of pseudouridine detection [179]. However, a weakness of all three of these methods is that any method involving an enrichment step cannot be truly quantitative. Another method, m^6A-level and isoform-characterization sequencing (m^6A-LAIC-seq), omits RNA fragmentation

and includes comprehensive RNA controls, thereby achieving a quantitative measurement of m^6A modification fraction on a transcript level [186]. However, by skipping the fragmentation step, this method sacrifices even the ∼100-nucleotide resolution of the original m^6A sequencing methods [60, 61]. Currently, the quantitative measurement of RNA modification fraction relies on low-throughput methods such as mass spectrometry [187] and SCARLET [72]. These methods have been important complementary approaches to sequencing methods, allowing for the validation of RNA modification sites and the quantitative determination of modification fraction. Mass spectrometry methods have also been used to simultaneously detect several types of RNA modifications at once, including m^6A [187]. In addition, advances in bioinformatics have examined the context-dependent effect of RNA modifications on sequencing signatures and have facilitated the detection of Watson–Crick-face RNA modifications in high-throughput sequencing data [188, 189].

Advances in sequencing methods and bioinformatic approaches have greatly improved the sensitivity, specificity, and resolution of RNA modification site mapping. However, high-throughput methods to determine modification fraction are lacking, and antibody artifacts, as well as the poor sensitivity and specificity of chemical methods, remain major problems in the detection of RNA modifications, including m^6A and pseudouridine. So far, all high-throughput methods for mapping m^6A rely on antibody enrichment, while all high-throughput methods for mapping pseudouridine rely on CMC treatment. The development of both antibody-based and chemical-based methods for the same RNA modification would allow for the combined application of different methods and thereby improve both sensitivity and specificity. Better yet, high-throughput methods that do not rely on either enrichment or chemical treatment could potentially reduce artifacts and improve the sensitivity and specificity of detection. Although advances in bioinformatics have facilitated the detection of Watson–Crick-face modifications in untreated RNA [182, 190], it remains unclear how such methods can be developed for non-Watson–Crick-face modifications such as m^6A and pseudouridine.

1.4 RNA-Binding Proteins

mRNAs interact with proteins throughout their life cycle. These interactions constitute a crucial and widespread mechanism for the regulation of mRNA transcription, processing, translation, and stability. RNA-binding proteins were traditionally identified and categorized based on their canonical RNA-binding domains [191]. However, the development of high-throughput technologies led to the identification of many more RNA-binding proteins. Interactome studies have expanded the number of RNA-binding proteins identified in the human genome to ∼1 500, corresponding to ∼7.5% of all human protein-coding genes [133, 192, 193]. Many of the newly discovered RNA-binding proteins in these studies do not possess canonical RNA binding domains. Furthermore, it has become increasingly clear that intrinsically disordered regions and low-complexity sequences have widespread roles in RNA binding [193].

1.4.1 Canonical RNA-binding domains

RNA-binding proteins are structurally diverse, containing ∼600 types of RNA-binding domains (RBDs) [133]. The vast majority of RBDs occur in only one or two RNA-binding proteins, and only twenty RBDs are found in more than ten RNA-binding proteins [133]. The most abundant RBDs include RNA recognition motifs (RRMs), hnRNPK-homology (KH) domains, double-stranded RNA-binding domains (dsRBDs), and zinc-finger (ZnF) domains. These RBDs are small domains of <100 amino acids and are composed of a combination of α-helices and β-strands.

RRMs and KH domains usually bind specifically to short sequence motifs in single-stranded RNA. The RRM domain is composed of a four-stranded antiparallel β-sheet and two α-helices [148]. In most cases, RNA binding occurs on the surface of the β-sheet through positively charged residues that form contacts with the phosphodiester backbone and aromatic residues that stack with the nucleotide bases. In addition, loops between the β-strands

or α-helices of the RRM often contribute to RNA binding. A single RRM domain can bind to 2–8 nucleotides in a sequence-specific manner. In contrast, KH domains bind to RNA in a hydrophobic cleft with additional specific hydrogen-bonding or electrostatic interactions [194]. Unlike RRMs, KH domains do not use aromatic residues to form stacking interactions with nucleotide bases. Moreover, KH domains vary widely in sequence specificity, but the RNA-binding cleft generally fits only four nucleotides.

In contrast to these single-stranded RNA-binding domains, dsRBDs bind specifically to double-stranded RNA by recognizing $2'$-hydroxyl groups and the width of the major groove in the A-form RNA helix [195]. An N-terminal α-helix in some dsRBDs can recognize distortions in the A-form helix, such as loops or bulges. In addition, certain dsRBDs bind double-stranded RNA in a sequence-specific manner through direct hydrogen-bonding contacts in the minor groove. Zinc-finger domains can bind to either single-stranded or double-stranded RNA in both a structure- and sequence-specific manner [196]. RNA recognition is mediated by a combination of hydrogen-bonding and stacking interactions, and contacts with double-stranded RNA occur in the major groove.

In addition to their function in RNA binding, many RBDs can bind DNA. Some RRM and KH domains that lack contacts with $2'$-hydroxyl groups have been shown to bind the same sequences in both RNA and DNA [194, 197]. In addition, the RRM2 domain of trans-activating response element (TAR) DNA-binding protein 43 (TDP43) can bind both RNA and DNA despite making direct contacts with a $2'$-hydroxyl group during RNA binding. When TDP43 binds to DNA, the residues that contact the $2'$-hydroxyl group in RNA instead form contacts with backbone phosphates in DNA [197]. Zinc-finger domains are found in classical DNA-binding transcription factors that bind to specific sequences through contacts in the major groove of double-stranded DNA [196]. In some cases, a single zinc-finger domain can use the same residues to recognize either DNA or RNA. In contrast, dsRBDs bind specifically to double-stranded RNA, as opposed to double-stranded DNA, RNA–DNA hybrids, or single-stranded RNA [195]. The competitive binding of DNA and RNA to the

same domain plays a role in the regulation of transcription factors by certain lncRNAs. For example, the lncRNA growth arrest-specific 5 (*GAS5*) inhibits DNA binding by the glucocorticoid receptor [197].

Many RNA-binding proteins contain multiple different types of RBDs or multiple repeats of the same class of RBD [133]. Combining multiple RBDs can increase the affinity and specificity of RNA binding by creating a multivalent interaction surface [191]. A modular structure also allows RNA-binding proteins to simultaneously bind to multiple distant sites in a single RNA or to multiple distinct RNA transcripts [191]. The RBDs of a modular RNA-binding protein are connected by linkers, which are usually intrinsically disordered but can often fold upon RNA binding. Interdomain linkers function to position RBDs relative to one another, which can be important for the binding specificity and function of the modular RNA-binding protein [191]. In addition, some linkers participate in RNA binding, thus impacting binding affinity and specificity. RBDs in modular RNA-binding proteins can also interact with one another through protein-protein interactions. These interactions can position the RBDs in particular orientations, such as side-by-side to create a continuous RNA-binding surface, or facing opposite directions to create RNA loops [148, 191]. RBDs can also form intra- or inter-molecular interactions with other types of domains, which can enhance or inhibit RNA binding, or alter sequence specificity [148, 191].

1.4.2 Non-canonical RNA-binding domains

The cataloguing of human RNA-binding proteins through mRNA interactome and other high-throughput studies revealed that about half of all RNA-binding proteins lacked any domain that had previously been shown to bind RNA [192, 198]. Furthermore, RNA-binding proteins are enriched in both intrinsically disordered regions and low-complexity regions compared to the overall human proteome [198].

Intrinsically disordered regions (IDRs) are protein regions that lack a defined structure, yet are functional in the cell [199]. IDRs lack bulky hydrophobic residues that would drive

the formation of a hydrophobic core, and their amino acid composition is frequently biased toward either polar or charged residues. IDRs can be just as compact as globular domains, but they are considered to be disordered due to their heterogeneous and often dynamic conformations. Short disordered regions exist as linkers between domains and are often found as conserved loops within DNA-, RNA-, or protein-binding domains. Longer disordered regions are also common: 44% of human proteins contain disordered regions of >30 amino acids, and 24% of human proteins have >30% of their residues in predicted disordered regions [199]. In contrast, IDRs are relatively rare in eubacteria and archaebacteria, where <5% of proteins are predicted to contain disordered regions of >30 residues in length [199]. In humans, IDRs of >500 residues are most commonly found in proteins with transcription-related functions. More generally, proteins enriched in IDRs frequently function as protein–protein interaction hubs, defined as proteins that interact with at least fifty proteins [192]. RNA-binding proteins and low-complexity regions are also enriched among hub proteins [192, 200].

Low-complexity regions (LCRs) are protein regions with an amino acid composition of low diversity. LCRs are found in about half of all proteins, which have an average of 25% of their residues in low-complexity sequences [201]. The amino acid composition of LCRs ranges from a single amino acid to several different residues, and their sequence pattern can be clustered, irregularly spaced, or periodic [200]. LCRs are typically disordered but can also be structured, depending on their amino acid composition [202]. LCRs are enriched in RNA-binding proteins and bind to about half of RNA-binding sites in cultured human cells [157]. The amino acid composition of LCRs in RNA-binding proteins differs from that of LCRs proteome-wide, instead showing an enrichment in residues that are often found at the interacting surfaces of canonical RBDs: the disorder-promoting residues serine (S), proline (P), and glycine (G); the positively charged arginine (R) and lysine (K); and tyrosine (Y) [198, 203, 204]. Besides tyrosine, the LCRs in RNA-binding proteins lack hydrophobic residues that promote protein folding. In addition to a bias in composition, certain

combinations of amino acids occur frequently, such as RGG, YGG, SR, and KK motifs [157].

The amino acid sequence RGG frequently occurs in clustered repeats and was previously observed to form a pattern called the RGG box [205]. The tri-RGG motif is similar to an RGG box and is defined as three repeats of the amino acid sequence RGG with no more than four residues between any two repeats: RGG($X_{0–4}$)RGG($X_{0–4}$)RGG [206]. A total of 31 human proteins have a tri-RGG motif, including the m^6A methyltransferase METTL14 and the m^6A reader hnRNPG. In both METTL14 and hnRNPG, the tri-RGG motif has been shown to function in RNA binding [3, 207]. The di-RGG motif, RGG($X_{0–4}$)RGG, is found in an additional 88 human proteins, while the tri-RG motif, RG($X_{0–4}$)RG($X_{0–4}$)RG, is found in over 300 human proteins [206]. Both of these motifs occur in hnRNPG in a region distinct from its tri-RGG motif, and are also found in the fragile X mental retardation protein (FMRP). While the previous motifs often occur near other RG sequences, the di-RG motif, RG($X_{0–4}$)RG, often exists independently and is very prevalent, found in >1 700 human proteins [206].

RGG regions are among the most common RNA-binding domains, second only to RNA recognition motifs [133], and many RGG motifs are known to function in RNA–protein interactions [206]. Solution and crystal structures demonstrate how a tri-RG-containing peptide from FMRP binds to a guanine-rich RNA in a sequence- and structure-specific manner, using a combination of hydrogen bonding and shape complementarity [208, 209]. The short amino acid sequence RGGGGR adopts a turn conformation, making direct contacts in the major groove of the duplex and at the duplex–quadruplex junction. In particular, the two arginines in this sequence form hydrogen bonds with guanines in the major groove, and both of these arginines are essential for RNA binding by the FMRP peptide [208, 209]. More generally, RGG motifs tend to have degenerate specificity for RNA binding partners, but often have a tendency to preferentially bind structured RNAs rich in guanine and cytosine [210].

RGG repeats also function in protein–protein interactions [206]. Multiple proteins repress translation by interacting with the eukaryotic translation initiation factor 4G (eIF4G) in a

manner dependent on their RGG motifs [211]. Moreover, the arginines in RGG motifs can be post-translationally methylated, which can alter their interactions with either proteins or RNA [206, 212]. Notably, methylated arginines are recognized by Tudor domains [206, 213]. Protein–protein interactions mediated by RGG motifs and Tudor domains are known to function in the regulation of transcription and splicing. Moreover, the methylation of RGG motifs can alter the subcellular localization of RGG-containing proteins in both normal and diseased cells, with functions in both nucleo-cytoplasmic transport and stress granule localization [206, 213].

Among the m^6A reader proteins identified so far, several have been shown to bind m^6A-containing RNAs using globular YTH domains or RNA recognition motifs [118–123, 137]. Although some of these m^6A reader proteins contain extensive low-complexity sequences, these regions have not been implicated in the binding of m^6A-containing RNAs. Given their increasingly recognized roles in RNA biology, low-complexity regions likely contribute to the recognition of m^6A-containing RNAs by some m^6A readers. In Chapter 3, I will describe how the protein hnRNPG uses a tri-RGG motif in a low-complexity region to selectively bind to m^6A-containing RNAs [3].

1.4.3 Ribonucleoprotein assembly

Macromolecules in the cell are organized into both membrane-bound and membrane-less compartments [214]. Ribonucleoprotein assemblies that act as membrane-less compartments include stress granules and processing bodies (P-bodies) in the cytoplasm, as well as nuclear speckles and paraspeckles in the nucleus. Because they lack membranes, these compartments rely on intermolecular interactions to maintain their integrity. These interactions are frequently weak multivalent binding events mediated by IDRs, including homologous interactions between repetitive protein regions, or heterologous interactions between positively charged IDRs and negatively charged RNA [214]. The assembly of ribonucleoproteins can be triggered by changes in environmental conditions. For instance, over a hundred endogenous

proteins in budding yeast reversibly assemble into stress granules as an adaptive response to heat shock [215]. Additionally, changes in cellular concentrations of protein or RNA can stimulate granule assembly [193]. For instance, the release of mRNAs from ribosomes upon stress results in an increased concentration of free RNA in the cytoplasm, which has been proposed to drive the formation of stress granules through RNA–RNA and RNA–protein interactions [216].

Ribonucleoprotein assemblies in the cell can exhibit liquid-like or solid-like properties. Liquid-like behaviors such as fusion and wetting have been observed *in vivo* for many ribonucleoprotein assemblies, including P-bodies in budding yeast [217], germ cell granules in the roundworm *Caenorhabditis elegans* [218], and stress granules in cultured human cells [217, 219]. Fluorescence bleaching experiments have shown that proteins are dynamically exchanged between these liquid-like cellular assemblies and their surrounding environment, with fluorescence recovery half-times of under one minute [218, 220]. While many cellular assemblies behave like liquids, others have more solid-like properties [217]. Moreover, some cellular granules have heterogeneous substructures, with different parts of the granule having different material properties [220]. In addition, mutations associated with neurodegenerative diseases can alter the assembly properties of both proteins and RNAs. For instance, a mutation in the protein hnRNPA1 can induce the formation of fibrous assemblies [219], while repeat expansions in RNA can alter RNA–RNA interactions and induce the formation of solid-like RNA assemblies [221].

The assembly of ribonucleoproteins has been reproduced using purified proteins *in vitro*. *In vitro* protein or ribonucleoprotein assemblies can resemble liquid droplets [219], various gel-like states [217, 222–224], or fibrous assemblies [183, 219]. For some of these liquid- and solid-like states, low-complexity regions are both necessary and sufficient for *in vitro* assembly [219, 224]. However, for other RNA-binding proteins, low-complexity regions are dispensable for assembly, and instead function to modulate the tendency towards assembly, for instance by shifting the temperature required to induce protein assembly [222].

Low-complexity regions can be disordered both in solution and in phase-separated liquid droplets [225], while they form amyloid-like cross-β structures in some types of hydrogels [224]. In some cases, liquid droplets can become solid-like and fibrous over time, and this 'maturation' has been proposed as a mechanism underlying some neurodegenerative diseases [226]. On the other hand, the transition from phase-separated liquids to assemblies with gel-like properties can also occur rapidly, resulting in spherical gel-like structures [222]. *In vitro* ribonucleoprotein assemblies can recruit other proteins or RNAs into a phase-separated state through protein–protein and RNA–protein interactions, and these interactions can in turn influence ribonucleoprotein assembly [219, 223]. For instance, RNAs can promote the liquid–liquid phase separation of RNA-binding proteins or alter the material properties of phase-separated liquid droplets [219]. Moreover, different RNAs can have different effects on droplet properties, possibly due to differences in RNA length or in the spacing of protein binding sites along the RNA transcript [227].

Phase separation has been implicated in heterochromatin formation and transcriptional regulation, among many other functions. Heterochromatin protein 1α (HP1α) phase-separates in a DNA- and phosphorylation-dependent manner, forming droplets that selectively recruit and exclude certain protein components [228]. In fruit fly embryos, HP1α phase-separates into liquid-like assemblies, but these assemblies become less dynamic as heterochromatin domains form. It has been proposed that the maturation of HP1α foci plays an important role in the formation of a heterochromatin–euchromatin barrier [229]. The intrinsically disordered CTD of RNAPII can nucleate LCR assembly and partition into liquid droplets or hydrogels [183, 225, 230]. Moreover, these interactions can be modulated by phosphorylation of the CTD, which prevents the CTD from binding to hydrogels [230], and by RNA, which enhances the interaction of an RNA-binding protein with the CTD [183, 225]. The assembly of the CTD with low-complexity regions has been proposed to function in the regulation of transcription [230]. Moreover, phase separation has been proposed to underly transcriptional control by super-enhancers, which are regions with multiple enhancers that

effect strong transcriptional activation and are occupied by a high density of interacting factors [231]. The RNAPII CTD, DNA bound by transcription factors, histone proteins bound by chromatin readers, and RNA bound by RNA-binding proteins have all been proposed to contribute to the valency and stability of these phase-separated assemblies [231].

Chapter 2

m^6A Modification in a Long Noncoding RNA Hairpin Predisposes Its Conformation to Protein Binding

Acknowledgement: This chapter is derived from an article published in *Journal of Molecular Biology* by Elsevier [1]. The authors of that article were: Katherine I. Zhou, Marc Parisien, Qing Dai, Nian Liu, Luda Diatchenko, Joseph R. Sachleben, and Tao Pan. Author contributions: Conceptualization, K.I.Z. and T.P.; Methodology, K.I.Z. and T.P.; Software, M.P.; Formal Analysis, M.P.; Investigation, K.I.Z.; Resources, Q.D. (oligonucleotide synthesis) and J.R.S. (NMR); Writing – Original Draft, K.I.Z.; Writing – Review & Editing, K.I.Z., M.P., Q.D., N.L., and T.P.; Supervision, L.D. and T.P.

2.1 Introduction

Dynamic RNA structures have extensive roles in the function of structural and regulatory lncRNAs, and in the regulation of mRNA transcription, splicing, translation, and stability [232]. Thus, the effect of m^6A on lncRNA and mRNA structure has the potential to influence many cellular processes. *In vitro* studies with model m^6A duplexes have demonstrated that m^6A can either stabilize or destabilize RNA secondary structures depending on its position within or at the end of a duplex [139]. Further evidence suggests that m^6A influences RNA structure*in vivo*. Parallel analysis of RNA structure (PARS) showed that RRACH motifs containing m^6A have a different RNA structural profile than RRACH motifs lacking m^6A modification [139]. Moreover, structural probing in an *in vivo* click selective $2'$-hydroxyl acylation and profiling experiment (icSHAPE) revealed an METTL3-dependent enhancement in reactivity at m^6A sites [141]. The widespread influence of m^6A on RNA secondary structure in cells could potentially have important consequences for the processing, function, and fate of mRNAs and lncRNAs.

Since changes in RNA structure can affect diverse cellular processes, the influence of

m^6A on mRNA and lncRNA structure has the potential to be an important mechanism for m^6A function in the cell. Indeed, an m^6A site in the lncRNA metastasis-associated lung adenocarcinoma transcript 1 (*MALAT1*) was shown to induce a local change in structure that increases the accessibility of a U_5-tract for recognition and binding by heterogeneous nuclear ribonucleoprotein C (hnRNPC). This m^6A-dependent regulation of protein binding through a change in RNA structure, termed 'm^6A switch,' affects transcriptome-wide mRNA abundance and alternative splicing. The discovery that an m^6A switch regulates hnRNPC binding revealed that m^6A-induced changes in mRNA and lncRNA structure have functional effects *in vivo*.

In this study, we further characterized the m^6A-induced structural changes in a 32-nucleotide hairpin derived from the m^6A switch in the human lncRNA *MALAT1*. Nuclear magnetic resonance (NMR) revealed that while the methylated hairpin maintains its overall structure, m^6A affects the distances between protons in the hairpin region where m^6A is located. Förster resonance energy transfer (FRET) studies further demonstrated that m^6A alters the conformation of the *MALAT1* hairpin to become more similar to the hnRNPC-bound hairpin, whereas hnRNPC binding induces similar conformations of both the methylated and unmethylated hairpins. Comparing A and m^6A hairpins shows that m^6A modification predisposes the RNA conformation to resemble more closely its conformation in the RNA–hnRNPC complex. The m^6A-induced structural changes in the *MALAT1* hairpin can serve as a model for a large family of m^6A switches that mediate the influence of m^6A on cellular processes.

2.2 Results

In previous studies, hnRNPC was found to preferentially bind an m^6A-modified hairpin composed of nucleotides $2\,556-2\,587$ of the lncRNA *MALAT1*, with an ∼8-fold higher affinity for the methylated hairpin [119]. Since hnRNPC is known to recognize single-stranded U-tracts of at least 5 U's in length, it was hypothesized that methylation of A2577 destabilizes

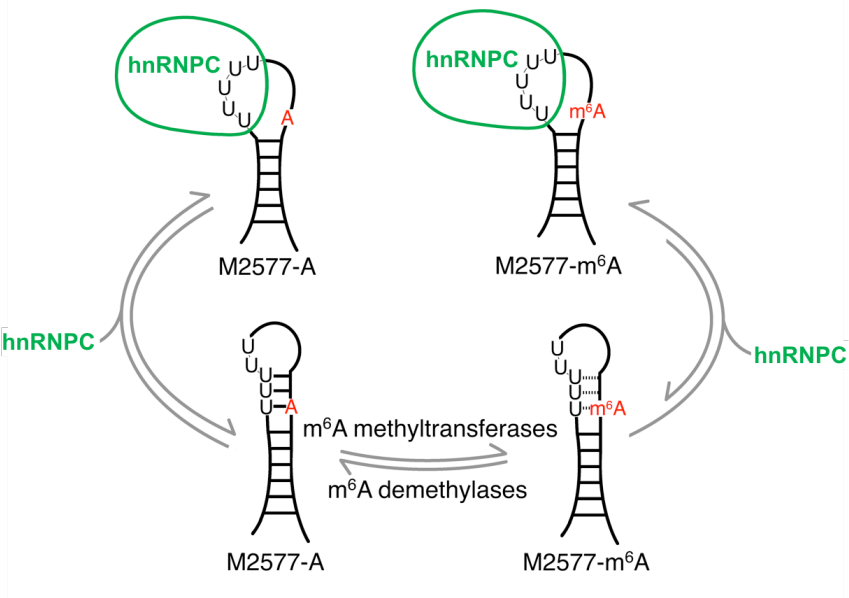


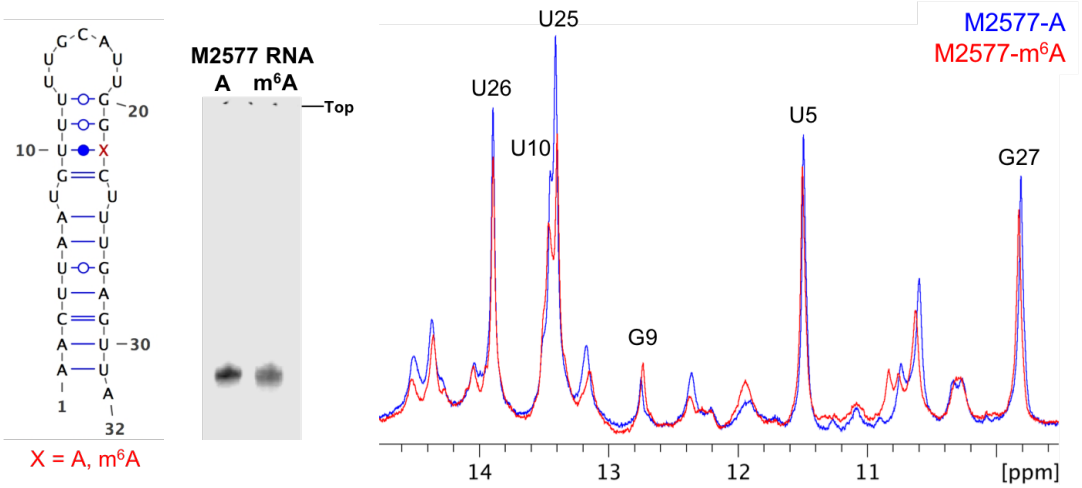
Figure 2.1: The m^{6}A-switch model. The human lncRNA *MALAT1* is reversibly methylated at position A2577. The protein hnRNPC binds to the U_5-tract in this hairpin from *MALAT1*, with an ∼8-fold higher affinity for the methylated hairpin. One of the U's in the hnRNPC binding site pairs with the methylation site A2577. The presence of m^{6}A weakens the base pair and increases the accessibility of the U-tract for protein binding.

the hairpin-stem, exposing the single-stranded U-tract for hnRNPC binding (Figure 2.1) [119, 142, 143]. Structural probing with RNase V1 and S1 nuclease was consistent with this m^{6}A-switch model, showing decreased stacking and increased single-strandedness in the region of the hairpin-stem surrounding A2577 upon m^{6}A modification [72, 119]. However, it was not known how extensive the global structural and dynamic differences are between the unmodified and m^{6}A-modified hairpins, and how m^{6}A modification enhances hnRNPC binding to the *MALAT1* hairpin. We address these questions here using NMR and FRET methods.

2.2.1 *NMR shows that methylation of the MALAT1 hairpin changes the conformation of a portion of the hairpin stem*

To examine the differences between the methylated and unmethylated *MALAT1* hairpin in solution, we collected 1D ^{1}H NMR spectra of both hairpins at 20 °C in 10% D_2O (Figure 2.2). Native gel electrophoresis demonstrated that the hairpins migrate as a single major species with the same mobility regardless of methylation status (Figure 2.2(b)). The 1D spectra of the two hairpins are largely similar, suggesting that the overall structure of the hairpin is maintained. In particular, the 9.5–14.8 ppm regions show that the chemical shifts of the imino protons H1 and H3 of G and U, respectively, are largely unaffected by methylation of the hairpin (Figure 2.2(c)).

We performed 2D nuclear Overhauser effect spectroscopy (NOESY) experiments of the



(a) (b) (c)

Figure 2.2: 1D NMR spectra show that the overall structure of the hairpin is maintained. (a) Secondary structure of the 32-nt M2577-A oligo from nucleotides 2 556 – 2 587 of *MALAT1*. The m^6A modification site (A22 in the oligo, or A2577 in *MALAT1*) is denoted with a red "X." The figure was made using Visualization Applet for RNA (VARNA) [233]. (b) 15% native PAGE of the unmethylated (M2577-A) and methylated (M2577-m^6A) hairpins in 25 mM Tris-acetate pH 7.4, 2.5 mM magnesium acetate. (c) Superimposed imino regions of the 1D ^{1}H NMR spectra of M2577-A (blue) and M2577-m^6A (red). Watergate solvent suppression 1D ^{1}H NMR spectra were measured under the conditions 1.12 mM RNA,10 mM Na_2HPO_4 pH 7.4, 2.5 mM $MgCl_2$, 90% H_2O / 10% D_2O (v/v), 20 °C.

Imino–imino pair	NOE intensity for M2577-A	NOE intensity for M2577-m^{6}A
G27H1–U5H3	1.00 ± 0.00	1.00 ± 0.00
U5H3–U26H3	0.20 ± 0.02	0.17 ± 0.03
U25H3–U26H3	0.16 ± 0.03	0.18 ± 0.01
G27H1–U26H3	0.06 ± 0.01	0.06 ± 0.00
G21H1–U11H3	0.05 ± 0.02	0.00 ± 0.01
G9H1–U10H3	0.04 ± 0.01	0.01 ± 0.00

Table 2.1: Imino–imino NOE intensity in 10% D_2O at 20 °C

methylated and unmethylated *MALAT1* hairpins at 4 °C and 20 °C in 10% D_2O to assign the imino protons and to detect differences in inter-proton distances (Table 2.1 and Figure 2.3(a)–2.3(b)). Sequential NOEs between imino protons of neighboring guanosines and uridines were used for imino proton assignments. Many of the same imino–imino NOEs were present in both methylated and unmethylated hairpins, suggesting that these base–base interactions are maintained and the overall structure of the hairpin does not change upon m^{6}A modification. However, two imino–imino NOEs were observed at 20 °C in the unmethylated hairpin, but not in the methylated hairpin: an NOE between the imino protons of U11 and G21, and an NOE between the imino protons of G9 and U10. Both NOEs involve bases within the U-tract that consists of the binding site for hnRNPC protein. The loss of these NOEs suggests a change in conformation in the upper part of the *MALAT1* hairpin-stem. Since NOEs are an indicator of through-space distance, where NOE signal falls rapidly with distance r as $1/r^6$, the loss of the NOE between U11 and G21 imino protons in particular is consistent with the model that this portion of the stem is less stably base paired in the methylated hairpin.

In addition, the methylated hairpin exhibited several changes in the amino–imino region of the NOESY spectrum (Figure 2.3(c)). The most pronounced changes were found in NOEs between amino region protons and the imino proton of U10. These amino region resonances likely correspond to protons from the A/m^{6}A22 that base pairs with U10. Similar to previous NMR studies with model m^{6}A duplexes, we observed NOEs of the m^{6}A22 H2 and H6 with the imino proton of the base-paired U10 [139]. Two NOEs were observed between the

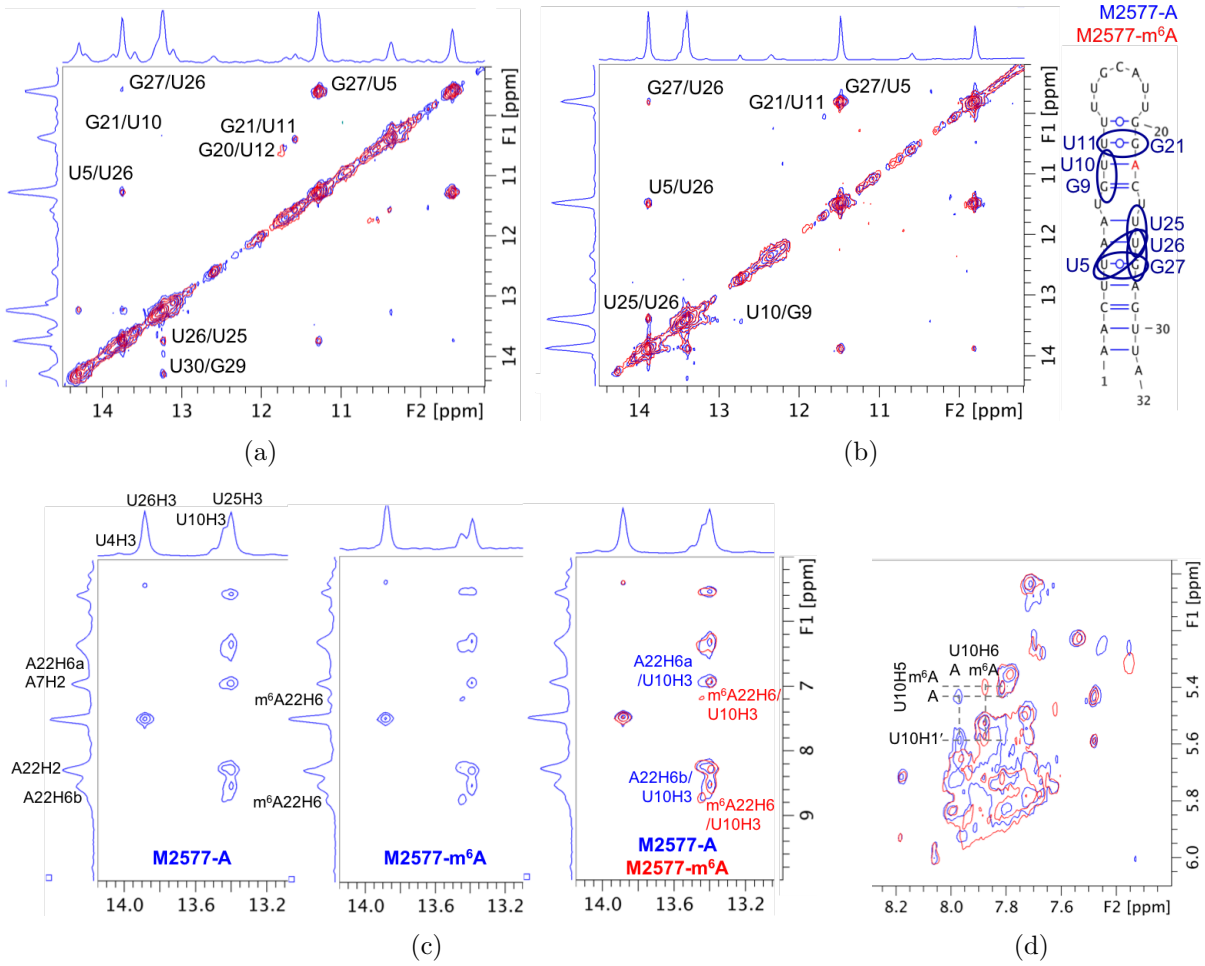


Figure 2.3: 2D NOESY spectra show that the upper stem is more dynamic in the methylated than in the unmethylated M2577 hairpin. (a) Superimposed imino regions of the 2D ^{1}H NOESY NMR spectra of M2577-A (blue) and M2577-m^{6}A (red) in 10% D_2O. The spectra of 0.47 mM RNA were measured at 4 °C with a 100 ms mixing time. (b) Superimposed imino regions of the 2D ^{1}H NOESY NMR spectra of M2577-A and M2577-m^{6}A in 10% D_2O. The spectra of 1.12 mM RNA were measured at 20 °C with a 100 ms mixing time. (c) Separate and superimposed amino–imino regions of the 2D ^{1}H NOESY NMR spectra of M2577-A and M2577-m^{6}A in 10% D_2O at 20 °C. (d) Superimposed H6/H8–H1′ regions of the 2D ^{1}H NOESY NMR spectra of M2577-A and M2577-m^{6}A in 100% D_2O. The spectra of 0.78 mM RNA were measured at 20 °C with a 100 ms mixing time.

m^6A22 H6 and the imino proton of U10, suggesting slow exchange between the *anti* and *syn* conformations of the N^6-methyl group. The NOEs of the U10 imino proton with the m^6A22 H6 proton were stronger than those with the A22 H6a and H6b protons of the unmethylated hairpin, likely due to slower rotation of the N^6-methylamino group, as has been previously proposed [139]. The NOE between the U10 imino and the A/m^6A22 H2 was equally intense in the methylated and unmethylated hairpins, suggesting that the hydrogen bond between the U10 imino proton and m^6A22 N1 is retained. Given that the single m^6A22 H6 proton shows two NOEs with the U10 imino proton, the m^6A–U base pair could be either singly or doubly hydrogen-bonded within the hairpin depending on the *syn* or *anti* conformation of the N^6-methyl group in m^6A. Previous studies with model m^6A duplexes found only one NOE between the m^6A H6 and the U imino [139]. This discrepancy is consistent with previous observations that the effect of m^6A on stability is strongly context dependent [139, 140]. The m^6A–U was two G–C pairs from the end of the model m^6A duplex used for NMR studies by Roost *et al.* [139], whereas in the *MALAT1* hairpin the m^6A–U is two G–U pairs from the loop, which could afford more flexibility for the N^6-methylamino group to rotate.

We further collected 2D NOESY spectra of the methylated and unmethylated *MALAT1* hairpins at 20 °C in 100% D_2O (Figure 2.3(d)). The resonances were broad and overlapping, such that the intra- and inter-nucleotide H6/H8–H1′ NOEs along the duplex could not be traced unambiguously. Nonetheless, a comparison between the NOESY spectra of the methylated and unmethylated hairpins showed that the H6/H8–H1′ regions were mostly similar, but with several distinct shifts in resonances. These shifted resonances most likely correspond to protons of the m^6A22–U10 base pair or of nearby nucleotides. Similar shifts in resonances have been observed in studies with model m^6A duplexes [139].

2.2.2 FRET shows that the methylated and unmethylated hairpins have different conformations

To further probe the influence of m^6A on the conformation of the *MALAT1* hairpin in solution, we designed two pairs of unmethylated and methylated *MALAT1* hairpins modified with a $5'$ indocarbocyanine-3 (Cy3) fluorophore and an internal fluorescein fluorophore in the hairpin stem (Figure 2.4(a)–2.4(b)). The constructs were named based on the position of the fluorescein fluorophore: the unmethylated and methylated hairpins Fl-8-A and Fl-8-m^6A contain fluorescein-dT at nucleotide position 8, while the unmethylated and methylated hairpins Fl-26-A and Fl-26-m^6A contain fluorescein-dT at nucleotide position 26.

We observed that m^6A modification resulted in a significant increase in FRET efficiency at ambient temperature (Figure 2.4(a)–2.4(c)). In contrast, the methylated and unmethylated hairpins had similar FRET efficiencies when denatured at 90 °C (Figure 2.4(d)). Based on these results, we suggest that m^6A increases the FRET efficiency by altering the conformation of the *MALAT1* hairpin, whereas m^6A does not alter the conformation of the unfolded oligo. Changes in FRET efficiency could be due to changes in the distance between donor and acceptor fluorophores or to changes in the relative orientation of the fluorophores. Since the fluorescein donor fluorophore was in the stem of the *MALAT1* hairpin, increased flexibility of the hairpin-stem upon m^6A modification could alter the position or orientation of the fluorescein fluorophore to increase the efficiency of energy transfer to Cy3 at the $5'$ end of the hairpin. This interpretation of the observed FRET efficiencies is consistent with the m^6A-switch model for the *MALAT1* hairpin, in which m^6A modification increases the flexibility of the hairpin-stem and exposes single-stranded RNA for protein binding.

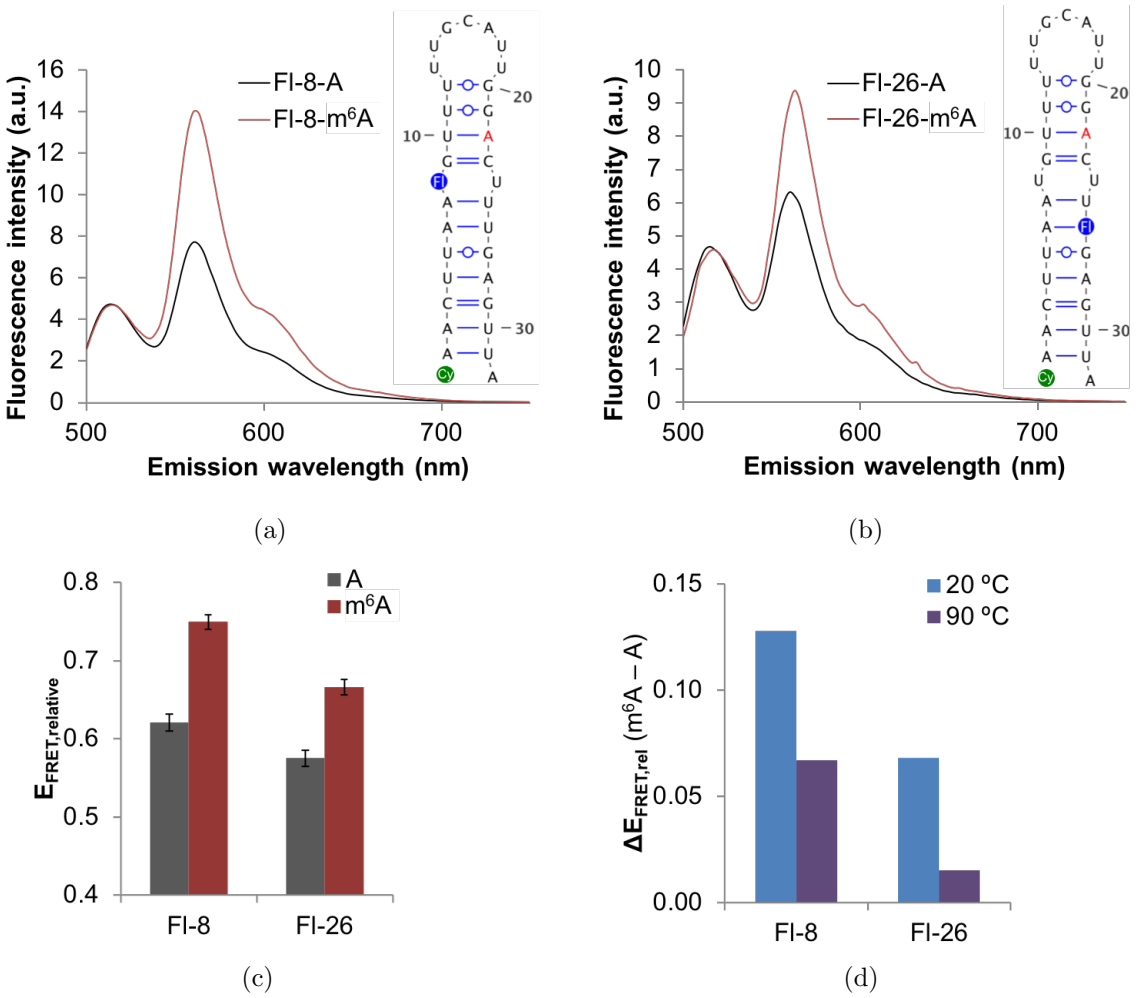


Figure 2.4: FRET shows that the methylated and unmethylated *MALAT1* hairpins have different conformations. (a) Fluorescence emission spectra of the FRET constructs Fl-8-A and Fl-8-m^{6}A upon excitation at 490 nm. Cy3 (green) is conjugated to the 5′ phosphate, and fluorescein-dT (Fl) is incorporated at the indicated position (blue) in each oligo. Spectra were measured under the conditions 500 nM RNA, 10 mM Tris pH 7.5, 100 mM KCl, 2.5 mM $MgCl_2$ at ambient temperature. Each spectrum is the average of 2–3 measurements. (b) Fluorescence emission spectra of Fl-26-A and Fl-26-m^{6}A upon excitation at 490 nm. (c) Relative FRET efficiencies ($E_{FRET,rel}$) of M2577-A and M2577-m^{6}A, calculated as $I_{563}/(I_{563}+I_{518})$, where I_x is the fluorescence emission intensity at x nm. FRET efficiencies are the mean of 6–8 measurements. Error bars represent ± one standard deviation. (d) Difference in the relative FRET efficiencies of M2577-A and M2577-m^{6}A at ambient temperature (20 °C) or at 90 °C.

2.2.3 FRET shows that the conformation of the methylated hairpin is more similar to the hnRNPC-bound RNA conformation

The above results demonstrate that m^6A modification of the *MALAT1* hairpin changes the conformation of the RNA alone. To evaluate the influence of m^6A modification on the conformation of the hnRNPC-bound hairpin, we added hnRNPC protein to the *MALAT1* hairpin constructs and measured the resulting FRET spectra. The FRET efficiencies of the methylated and unmethylated *MALAT1* hairpins became more similar upon addition of hnRNPC (Figure 2.5(a)–2.5(b)), whereas the FRET efficiencies did not change upon addition of Proteinase K as a control protein (Figure 2.5(c)). Thus, although m^6A modification alters the conformation of the *MALAT1* hairpin alone, in the ribonucleoprotein (RNP) complex the RNA has the same conformation regardless of modification status. We suggest that the difference in the affinity of hnRNPC for the methylated and unmethylated hairpins is the result of a difference in the conformation of the unbound RNA hairpins, whereas the ribonucleoproteins have the same conformation and energetics regardless of RNA methylation. The 8-fold difference in the K_d for hnRNPC binding can be accounted for by a 1.2 kcal/mol destabilization of the hairpin duplex by m^6A to expose the U-tract for hnRNPC binding, consistent with the previously observed 0.5–1.7 kcal/mol destabilizing effect of m^6A on model RNA duplexes [139]. The change in the stability of the RNA hairpin explains how formation of an RNP with the methylated hairpin is more thermodynamically favorable than formation of an RNP with the unmethylated hairpin, even though the methylated and unmethylated RNPs are similar in conformation and energy.

For both sets of constructs, the change in FRET efficiency upon protein binding was more drastic for the unmethylated hairpin than for the methylated hairpin (Figure 2.5(c)), suggesting that the conformation of the methylated hairpin is more similar to that of the hnRNPC-bound hairpin. Since the conformations of the free and bound m^6A-modified hairpin are similar, the conformational change in the conversion of the free form to the bound form might require less energy than the conversion of the unmodified hairpin from

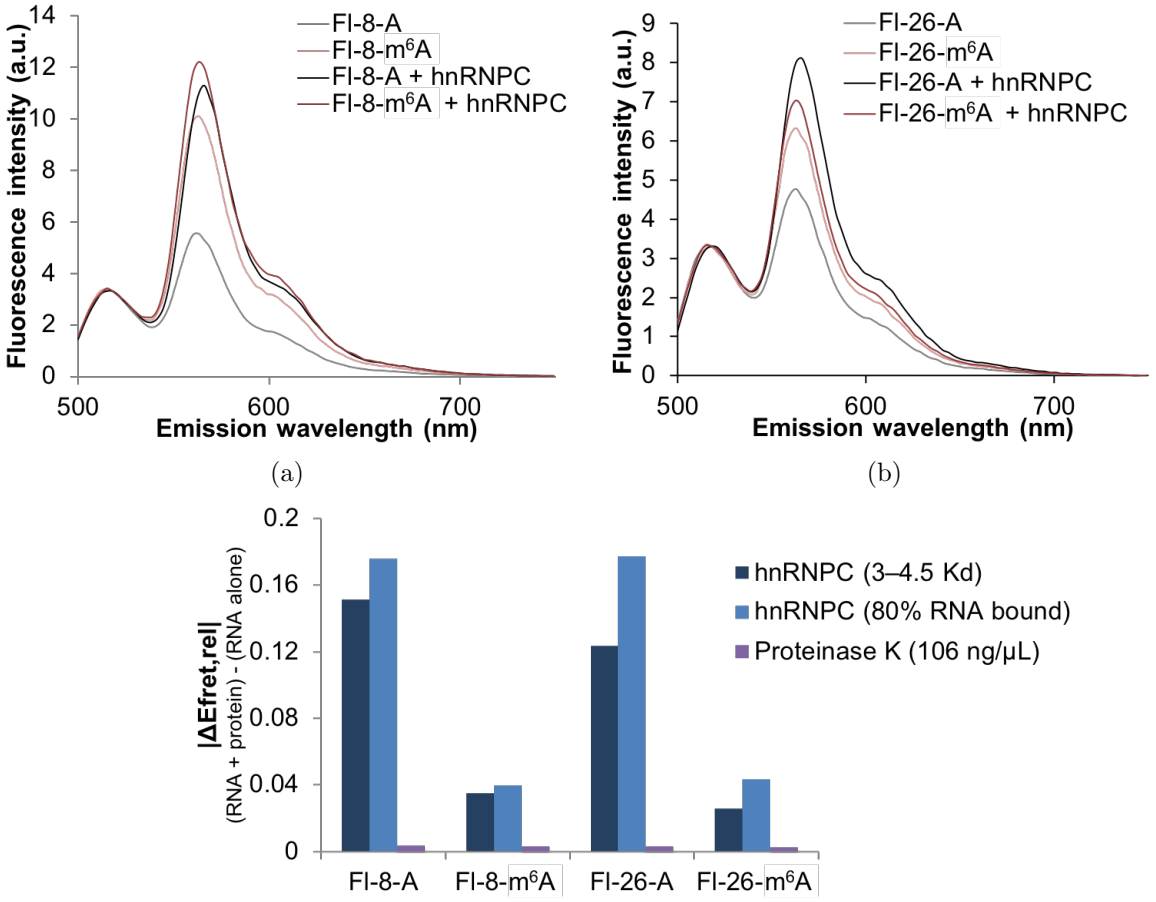


Figure 2.5: FRET of RNPs containing the M2577-A and M2577-m^{6}A hairpins. The RNPs show similar FRET, suggesting that the conformation of the RNA in the RNP is the same regardless of the presence of m^{6}A. In addition, the methylated hairpins exhibit a smaller change in FRET upon hnRNPC binding. (a) Fluorescence emission spectra of 500 nM Fl-8-A and Fl-8-m^{6}A with or without addition of hnRNPC at a concentration of 3–4.5·K_d (K_d = 722 nM for M2577-A, K_d = 93 nM for M2577-m^{6}A) [119]. Spectra were measured under the conditions 500 nM RNA, 10 mM Tris pH 7.5, 100 mM KCl, 2.5 mM $MgCl_2$ at ambient temperature at excitation wavelength 490 nm. (b) Fluorescence emission spectra of 500 nM Fl-26-A and Fl-26-m^{6}A with or without addition of hnRNPC at a concentration of 3–4.5·K_d. (c) Change in the relative FRET efficiency ($\Delta E_{FRET,rel}$) of each hairpin (500 nM) upon addition of: 3–4.5·K_d hnRNPC (2.17 μM hnRNPC for M2577-A; 410 nM hnRNPC for M2577-m^{6}A), hnRNPC such that $[RNP]/[RNA]_{total} = 80\%$ (3.29 μM hnRNPC for M2577-A; 770 nM hnRNPC for M2577-m^{6}A), or 106 ng/μL Proteinase K (equivalent to weight/volume concentration of 3.25 μM hnRNPC).

the free conformation to the bound conformation. In this manner, m^6A modification seems to set up the hairpin for hnRNPC binding by inducing a conformation more similar to the protein-bound form.

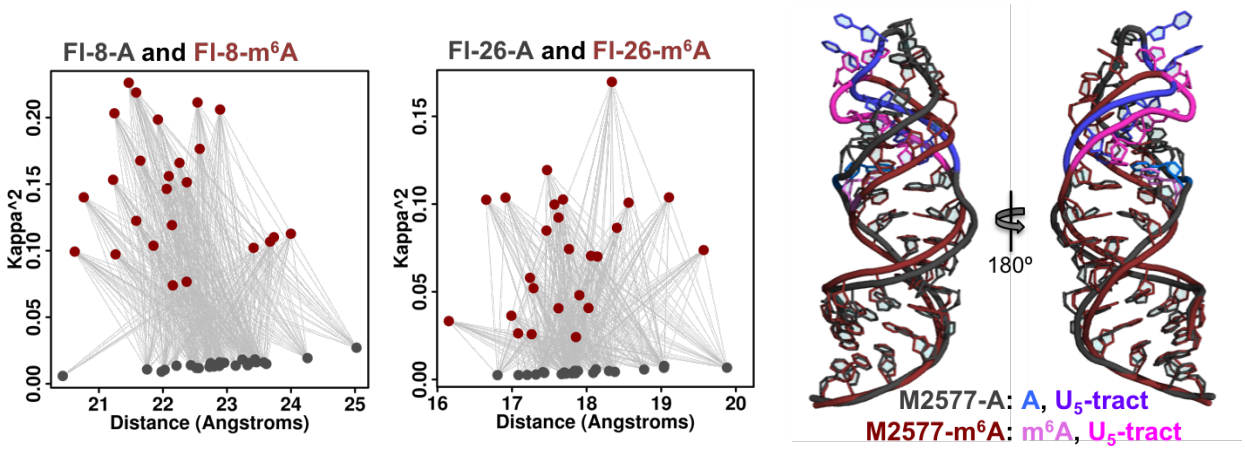
2.2.4 Structural modeling shows how m^6A can alter the conformation of the MALAT1 hairpin

Using the RNA tertiary structure prediction program MC-Sym [234], we generated 9 999 models for the *MALAT1* hairpin and selected models corresponding to the methylated and unmethylated hairpins based on four simultaneous criteria: (1) best fit to the FRET data, (2) best fit to the 2D NOESY data collected with 100 ms mixing time in 10% D_2O, (3) best P-Scores, and (4) maximization of relative FRET yields. While MC-Sym can use NMR data to guide model generation, FRET data involve pairs of models, so they are not amenable to interpretation during model generation. Instead of generating models based on the experimental data, we first generated a large pool of models, then chose those that satisfy all the data. True conformational sampling would require the use of molecular dynamics simulations, but due to very slow RNA dynamics, this approach is not attempted here.

The selected models were narrowed down to 25 models corresponding to the unmethylated hairpin and 25 models corresponding to the methylated hairpin. The parameters used to calculate the theoretical FRET efficiencies (the orientation parameter κ^2 and the distance between fluorophores) are plotted in Figure 2.6(a). While the distribution of distances is similar for both the "A" and "m^6A" sets, the models corresponding to the methylated hairpin ("m^6A" set) show more variation in the orientation parameter κ^2. Since the position of the Cy3 fluorophore was kept invariant in all 9 999 models, κ^2 depends primarily on the position and orientation of the fluorescein fluorophore in the hairpin stem. The observation that the "m^6A" set shows more variation in κ^2 implies that a wider range of different fluorescein fluorophore orientations is consistent with the FRET and NMR data, raising the possibility that the bases in the stem of the methylated hairpin have greater dynamic flexibility or can

adopt multiple distinct conformations.

The centroid of the 25 models was used to generate a single model each for methylated and unmethylated *MALAT1* hairpins (Figure 2.6(b)). The superimposed model structures reveal an m^{6}A methylation-dependent change in the conformation of the upper stem and loop of the *MALAT1* hairpin, including the backbone and nucleobases of the U-tract bound by hnRNPC. Thus, m^{6}A modification of the hairpin induces a conformational change that propagates through the hairpin structure sufficiently to influence the structure of the hnRNPC binding site, which supports the model that the effect of m^{6}A on the hairpin structure indirectly causes a change in the hnRNPC binding affinity.



(a) (b)

Figure 2.6: Structural models for M2577-A and M2577-m^{6}A based on FRET and NMR data. (a) Plot of 25 selected structures for M2577-A (gray) and M2577-m^{6}A (dark red), in terms of κ^2 and distance between fluorophores for Fl-8-A/m^{6}A and Fl-26-A/m^{6}A. The structures were selected from an initial set of 9 999 tertiary structures for the M2577 hairpin [234] (c) Structural models of M2577-A (gray) and M2577-m^{6}A (dark red), computed as the centroid of the 25 selected structures. The m^{6}A modification site and the U$_5$-tract are highlighted in shades of blue for the unmethylated *MALAT1* hairpin, and in magenta for the methylated hairpin.

2.3 Discussion

In this study, we used biophysical methods and modeling to examine the effect of m^6A modification on the *MALAT1* hairpin. Our NMR and FRET results demonstrate that the general structure of the *MALAT1* hairpin is maintained, but the nucleobases of the hairpin-stem are more flexible and solvent accessible upon m^6A modification. These results support the m^6A-switch model, in which m^6A regulates protein binding through its influence on RNA structure [119].

While previous studies examined the influence of m^6A on the structure and stability of model RNA duplexes, no past studies used NMR to investigate a physiological m^6A-modified RNA [139, 140]. The study of nucleic acids by NMR is already challenging due to low proton density and high spectral overlap [235, 236]. The terminal loop, internal loop, and noncanonical G–U pairs of the *MALAT1* hairpin further complicate the detection and assignment of imino protons by reducing the number of detectable protons, interrupting the continuity of the stem, and introducing ambiguity in the assignment of imino–imino NOEs. The possibility of dynamic changes in structure, base pairing, and oligomerization state introduces additional difficulties in the study of a naturally occurring RNA hairpin. In the future, structural studies of physiological m^6A-modified RNA might take advantage of selective labeling with ^{15}N or ^{13}C isotopes. Such methods would enable direct observation of hydrogen bonding, unambiguous identification of noncanonical base–base interactions, and better resolution of local changes in conformation [235, 236].

The FRET constructs used in this study showed that m^6A modification influences the observed FRET efficiency, likely due to an m^6A-induced change in the conformation of the *MALAT1* hairpin. However, ensemble FRET studies cannot distinguish a homogeneous population adopting a single conformation from a heterogeneous population with multiple subpopulations or with dynamically changing conformations [237]. It is very possible that the m^6A-modified hairpins not only have a different average structure, but are also more dynamic or adopt a more heterogeneous set of different conformations. It would be very interesting

to study our constructs using single-molecule FRET in order to better understand how m^6A influences the conformational dynamics of the *MALAT1* hairpin. Single-molecule FRET studies are well-suited to studying dynamic systems and have provided insight into processes such as RNA folding and RNP formation [238]. In addition to clarifying the structure and dynamics of the *MALAT1* hairpin, single-molecule FRET could further elucidate how m^6A affects hairpin folding and protein binding.

The *MALAT1* hairpin is the first identified example of an m^6A switch, but the changes induced by m^6A modification of the *MALAT1* hairpin are likely generalizable to a much larger family of m^6A-regulated RNA structures. Over 2 000 high-confidence m^6A switches have been identified at hnRNPC binding sites [119]. In addition, the m^6A-switch mechanism has the potential to regulate the binding of other RNA-binding proteins through altered accessibility of their single-stranded RNA binding motifs or through changes in their cognate RNA structures. Thus far, only proteins in the YTH family are known to directly bind m^6A through a well-defined structural mechanism [127]. The m^6A-switch mechanism expands the pool of candidate m^6A readers to a much wider array of RNA-binding proteins. Indirect m^6A readers might be pervasive but difficult to discover because in many cases only a subset of their targets might be regulated by m^6A modification. For example, m^6A switches seem to regulate $\sim$8% or 40 000 of all known hnRNPC binding sites [119]. Moving forward, it will be important to investigate other indirect m^6A readers and the mechanisms by which m^6A alters RNA structure to influence protein binding.

As the most abundant post-transcriptional modification in eukaryotic mRNA and lncRNA, m^6A could have pervasive regulatory roles in the regulation of mRNA transcription, splicing, and translation, and in influencing the structure and function of lncRNAs. In addition, m^6A modification might influence RNA structures in other classes of noncoding RNA. For example, m^6A methylation of primary microRNAs (pri-miRNAs) has been shown to be crucial for recognition by the microprocessor complex, though it is unclear in this case whether m^6A functions by influencing RNA structure or through direct recognition [153]. While m^6A mod-

ification likely regulates many m^6A switches using the same mechanism as in the *MALAT1* hairpin, m^6A could potentially use other mechanisms to regulate RNA structures such as disrupting a tertiary hydrogen bond [70]. Even in an RNA stem-loop, the influence of m^6A on RNA structure is dependent on context, as m^6A can either stabilize or destabilize depending on its position. It will be important to investigate the diverse and context-dependent effects of m^6A on RNA structure and dynamics, and how these are linked to the functions of m^6A in the cell. As the first example of an m^6A-induced structural change in a cellular RNA, the *MALAT1* m^6A switch is an initial model for a potentially much more general mechanism by which m^6A achieves its functions in the cell.

2.4 Materials and Methods

2.4.1 RNA synthesis and purification

RNA oligos containing two fluorophore modifications in each sequence were synthesized by Expedite DNA synthesizer on a 1 μmol scale. Cy3 phosphoramidite and fluorescein-dT phosphoramidite were purchased from Glen Research. m^6A phosphoramidite was prepared by following our reported procedure [239]. All the other phosphoramidites and beads were purchased from Chemgene. After oligo synthesis, the RNA oligos were first deprotected by treatment with 30% ammonium hydroxide and ethanol (3:1, v/v) at 55 °C for 4 h. Once cooled to ambient temperature, the supernatant was dried in a SpeedVac and the resulting pellets were further deprotected by treatment with a mixture of dimethyl sulfoxide (100 μL) and hydrogen fluoride triethylamine (125 μL) at 65 °C for 2.5 h. After cooling to ambient temperature, 22.5 μL sodium acetate (3 M) and *n*-butanol (1 mL) were added, and the mixture was vortexed and precipitated at −80 °C for 1 h. After centrifugation, the supernatant was removed, and the pellets were washed with 70% ethanol and purified on an 8% acrylamide:bisacrylamide (29:1), 7 M urea, 89 mM Tris-borate (pH 8.3), 2 mM Na_2EDTA (ethylenediamine-tetraacetic acid) gel. RNA was excised from the gel by UV shadowing and

eluted in 50 mM potassium acetate, 200 mM KCl (pH 7.5) by the crush-and-soak method. Eluted RNA was precipitated in ethanol, then resuspended and stored in H_2O at -20 °C.

RNA oligos M2577-A and M2577-m^{6}A were synthesized, deprotected, and purified in a similar way except that we used a Mermade synthesizer on a 5 μmol scale.

2.4.2 hnRNPC protein expression and purification

Rosetta BL21 *Escherichia coli* were transformed with a pGEX-6p-1 plasmid containing the full-length hnRNPC1 coding sequence inserted between the BamHI and XhoI restriction sites. The transformed bacteria were grown to saturation at 37 °C, 200 rpm in Luria–Bertani Lennox media with 100 μg/mL ampicillin and 50 μg/mL chloramphenicol, then diluted 1:100, grown in the same culture media to an absorbance of 0.6 at 600 nm, and induced with 2.5 mM isopropyl β-D-1-thiogalactoside (IPTG). The bacteria were grown an additional 5 hours at 37 °C, 200 rpm, then harvested and sonicated at 4 °C. GST–hnRNPC1 fusion protein was isolated from the soluble lysate using GST-Bind resin (Novagen), and then cleaved by GST-tagged PreScission Protease for 16 hours at 4 °C. The purified full-length hnRNPC1 protein was stored in 10 mM Tris (pH 7.5), 100 mM KCl, 2.5 mM $MgCl_2$, 30% glycerol (v/v) at -80 °C.

2.4.3 NMR spectroscopy

NMR data were acquired on a Bruker AVANCE III 600 MHz (14 Tesla) NMR spectrometer with a 5 mm pulsed field gradient (z-axis) triple HCN probe, and were processed using TopSpin v3.2 software. All NMR experiments were conducted at 20 °C, with trimethylsilyl propanoic acid (TSP) as the ^{1}H chemical shift reference. Gel-purified RNA in H_2O was centrifuged 10 minutes at 17 K $\cdot$ g to sediment any particulate matter. The supernatant RNA was combined with Na_2HPO_4 buffer (pH 7.4) and incubated 1 minute at 90 °C, then 3 minutes at ambient temperature. $MgCl_2$, D_2O, and TSP were added to a final volume of 500 μL with 10 mM Na_2HPO_4 pH 7.4, 2.5 mM $MgCl_2$, 90% H_2O / 10% D_2O (v/v).

The samples were then incubated 5 minutes at ambient temperature and stored at 4 °C until data collection. 1D ^{1}H NMR spectra of the RNA hairpins were collected at 1.12 mM concentration, with 1 028 scans. 2D ^{1}H NOESY spectra in 90% H_2O / 10% D_2O (v/v) were acquired with 100 ms mixing time, with 256 scans. 2 048 points were taken in F2 and 512 points in F1, with a recycle delay of 1 second and a spectral width of 22 ppm in both dimensions. The RNA concentration was 1.12 mM for the 2D NOESY scans at 20 °C, and 0.47 mM for the 2D NOESY scans at 4 °C. 2D ^{1}H NOESY spectra of 0.78 mM RNA in 100% D_2O were acquired with 100 ms mixing time, with 256 scans. 2 048 points were taken in F2 and 512 points in F1, with a recycle delay of 1 second and a spectral width of 9 ppm in both dimensions.

2.4.4 FRET experiments

FRET spectra were acquired on a HORIBA FluoroLog-3 Spectrofluorometer equipped with a Peltier controller, and processed using FluorEssence v3.5 software. 1 μM gel-purified RNA in H_2O was combined with Tris buffer (pH 7.5) and incubated 2 minutes at 90 °C, then 3 minutes at ambient temperature. KCl and $MgCl_2$ were added to a final volume of 100 μL with conditions 500 nM RNA, 10 mM Tris (pH 7.5), 100 mM KCl, 2.5 mM $MgCl_2$. For experiments with protein binding, hnRNPC was added with the same final buffer conditions. For experiments with denatured RNA, the sample was incubated at least 5 minutes at 90 °C, and the spectra were measured with the Peltier controller set at 90 °C. The samples were transferred to the cuvette and emission spectra were collected from 500 nm to 750 nm using the excitation wavelength 490 nm, with excitation and emission spectral slit widths of 2 nm and 5 nm, respectively. A buffer solution of 10 mM Tris (pH 7.5), 100 mM KCl, 2.5 mM $MgCl_2$ was used as the emission spectrum blank. FRET efficiencies were calculated as $E_{FRET,relative} = I_A/(I_D + I_A)$, where I_D is the emission intensity at 518 nm and I_A is the emission intensity at 563 nm.

2.4.5 Structural modeling

The 33-nucleotide sequence 5′-UAACUUAAUGUUUUUGCAUUGGACUUUGAGUUA with secondary structure "(((((((((.((((.....)))).)))))))))", where parentheses denote base pairs and dots denote non-base-paired residues, was used to generate 910 decoy RNA tertiary structures. The decoys varied from one another only in the U0–A32 base pair, where the 5′-most nucleobase U0 was added as a placeholder for the Cy3 fluorophore present in the FRET oligos. Each of the 910 decoys was used to generate 9 999 RNA tertiary structure models for the *MALAT1* hairpin using the MC-Sym computer program. Within each decoy set, the U0–A32 and A1–U31 base pairs were invariant. The decoy set that generated the most pairs that fit the FRET data for either Fl-8-A/m^{6}A or Fl-26-A/m^{6}A was used to select models for the methylated and unmethylated hairpins. Rather than assigning weights to the various experimental parameters, we used the experimental data as filters in a sequential fashion, and the final selected models do not depend on the order of application of the filters.

Models were selected from the 9 999 structural models in the decoy set based on four simultaneous criteria: (1) best fit to the FRET data, (2) best fit to the 2D NOESY data, (3) best P-Scores, and (4) maximization of relative FRET yields.

(1) *Best fit to the FRET data.* The theoretical FRET efficiencies were calculated as

$$E_{\mathrm{FRET,rel}} = \frac{1}{1+(R^6/{R_\mathrm{o}}^6)}$$

with

$${R_\mathrm{o}}^6 = (55.7\mathring{A})^6 \cdot \kappa^2 \cdot \tfrac{3}{2}$$

and

$$\kappa = \mathbf{D}\cdot\mathbf{A} - 3\,(\mathbf{R}\cdot\mathbf{D})\,(\mathbf{R}\cdot\mathbf{A})$$

where **D** and **A** are the unit vectors oriented from N1 to C4 of the uridine nucleotides corresponding to the donor and acceptor fluorophores, respectively, **R** is the unit vector

oriented from the donor position H3 to the acceptor position H3, and R is the distance from the donor H3 to the acceptor H3. Only pairs of structures for which the theoretical E_{FRET}(A) / E_{FRET}(m^6A) ratios were within 0.01 of the experimental ratios for Fl-8-A/m^6A and Fl-26-A/m^6A were kept (122 844 A-m^6A pairs).

(2) *Best fit to the 2D NOESY data.* To extract inter-proton distance information from the 2D NOESY data at 20 °C in 10% D_2O with 100 ms mixing time, we assumed a linear relationship between peak intensity and mixing time:

$$\eta = 2W_o\ t_{mix}$$

where η is the NOE peak intensity, W_o is the rate of the zero-quantum transition, and t_{mix} is the mixing time (100 ms). Using this approximation, the inter-proton distance r is related to the peak intensity by

$$r^6 \propto \frac{1}{W_o} \propto \frac{1}{\eta}$$

so the experimental NOE intensities and the relative distances between imino protons in the modeled tertiary structures were used to calculate $log({r_x}^6/{r_o}^6)$ ratios, where r_x is the inter-proton distance for a pair of imino protons, and r_o is the distance between G27 H1 and U5 H3. The least squares differences between the five experimental and modeled $log({r_x}^6/{r_o}^6)$ ratios were then used to classify the modeled structures as either "A" or "m^6A" depending on whether they were a closer fit to the unmethylated or methylated hairpin, respectively. Using this method, 8 176 structures were assigned as "A," while 1 823 were assigned as "m^6A." Only pairs of structures for which the "A" and "m^6A" assignments were consistent with the assignments based on the FRET efficiencies were kept.

(3) *The P-Score* for each modeled RNA tertiary structure was calculated based on the phosphate chain torsion angles in the predicted tertiary structures as described previously [240]. P-scores involve as many as four consecutive phosphate groups, and their aim is to assess how natural the modeled RNA looks like given the backbone trace. Only the top 5 000 of the 9 999 tertiary structures in the decoy set were kept.

(4) *Maximization of relative FRET yields.* Once the original 9 999 structures in the decoy set were filtered based on their FRET fit, NOE fit, and P-Scores, there were 276 remaining structures corresponding to the unmethylated hairpin, and 713 structures corresponding to the methylated hairpin. These were narrowed down to 25 models each corresponding to the unmethylated and methylated hairpins by maximizing the density of A–m^{6}A pairs. To achieve density maximization, a structural model that has been selected to be a representative of the "A" state (Figure 2.6(a), gray dots) must maximize the number of structural models in the "m^{6}A" state (Figure 2.6(a), dark red dots) for which the relative FRET efficiency yielded is close to the one experimentally observed (gray lines connecting the dots). The same principle was applied while populating the "m^{6}A" state; models must maximize the number of "A" state relative FRET yields.

Chapter 3

m^6A Alters RNA Structure to Regulate Binding of a Low-Complexity Protein

Acknowledgement: This chapter is derived from an article published in *Nucleic Acids Research* by Oxford University Press [3]. The authors of that article were: Nian Liu*, Katherine I. Zhou*, Marc Parisien, Qing Dai, Luda Diatchenko, and Tao Pan (*equal contributions). Author contributions: Conceptualization, N.L., K.I.Z., and T.P.; Methodology, N.L., K.I.Z., and T.P.; Software, M.P.; Formal Analysis, M.P.; Investigation, N.L. (pull-down, gel shift, structural probing, and sequencing experiments) and K.I.Z. (cross-linking, electron microscopy, and pull-down experiments); Resources, Q.D. (oligonucleotide synthesis); Writing – Original Draft, N.L. and K.I.Z.; Writing – Review & Editing, N.L., K.I.Z., M.P., and T.P.; Supervision, L.D. and T.P.

3.1 Introduction

N^6-methyladenosine (m^6A) is the most abundant internal modification in eukaryotic mRNA, and affects almost every stage of the mRNA life cycle. The YTH domain-containing proteins can specifically recognize m^6A modification to control mRNA maturation, translation, and decay. m^6A can also alter RNA structures to affect RNA–protein interactions in cells. Here, we report that heterogeneous nuclear ribonucleoprotein G (hnRNPG) is a new m^6A reader protein that uses a low-complexity region to recognize a motif exposed by m^6A modification. We first identified hnRNPG as a protein that preferentially binds to the m^6A-modified form of a hairpin from the long noncoding RNA (lncRNA) metastasis-associated lung adenocarcinoma transcript 1 (*MALAT1*). hnRNPG binds a purine-rich region that can overlap with the m^6A consensus sequence and is exposed upon m^6A modification. Moreover, hnRNPG binding is mediated by a low-complexity region rather than a canonical RNA binding domain. Transcriptome-wide studies further identified 13 191 high-confidence m^6A sites bound

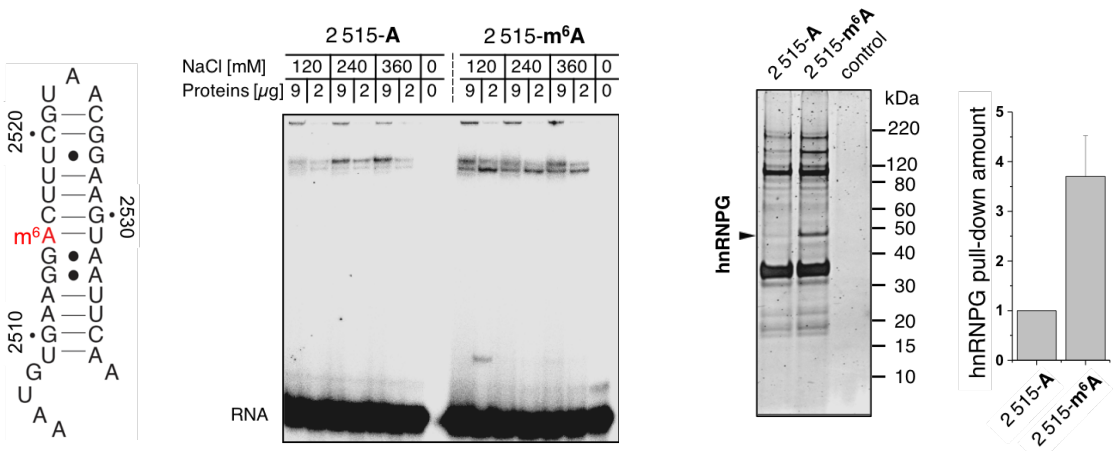
by hnRNPG, while *HNRNPG* knockdown and m^{6}A methyltransferase knockdown led to correlated changes in mRNA splicing. Thus, hnRNPG uses its low-complexity region to bind purine-rich sequences exposed upon m^{6}A modification of RNA, and thereby functions in the regulation of gene expression and alternative splicing. Low-complexity regions are pervasive among mRNA-binding proteins. Our results show that m^{6}A-dependent RNA structural alterations can promote direct binding of m^{6}A-modified RNAs to low-complexity regions in RNA-binding proteins.

3.2 Results

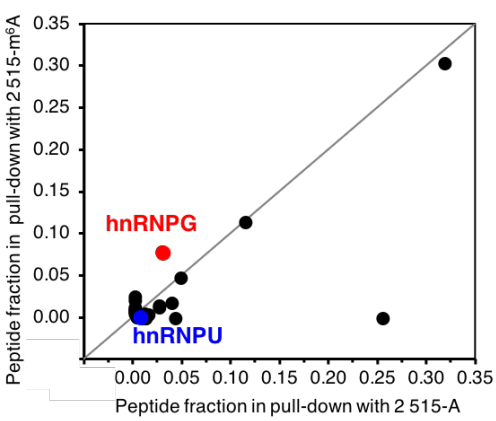
3.2.1 hnRNPG preferentially binds m^{6}A-modified RNA

In order to identify nuclear m^{6}A reader proteins, we conducted an RNA pull-down assay using methylated and unmethylated forms of a 34-nucleotide hairpin from the lncRNA *MALAT1* (Figure 3.1(a)). This hairpin contains a single m^{6}A site corresponding to position 2 515 of *MALAT1*, which is 63% m^{6}A-modified in HeLa cells [72]. Following incubation with HeLa nuclear extract, the methylated and unmethylated forms of the *MALAT1* hairpin pulled down different protein complexes, which were resolved on native gels (Figure 3.1(b)). Denaturing gel electrophoresis (Figure 3.1(c)) and mass spectrometry (Figure 3.1(d)–3.1(e)) further revealed that the protein hnRNPG was enriched in the pull-down with the m^{6}A-methylated hairpin. These results identify hnRNPG as an m^{6}A reader protein that preferentially binds an m^{6}A-modified hairpin from the lncRNA *MALAT1*.

hnRNPG is a ubiquitously expressed RNA binding protein encoded by the gene *RBMX* (RNA binding motif gene on X chromosome) [241]. A dominant function of hnRNPG is the regulation of alternative splicing [242, 243], but hnRNPG also functions in other cellular processes including DNA repair [244], sister chromatid cohesion [245], and transcriptional regulation [246]. Notably, hnRNPG has been shown to either positively or negatively regulate the inclusion of several disease-related exons [242, 243]. In addition, hnRNPG plays an



(a) (b) (c)



(d)

Gene Symbol	Accession
ACTB	gil28336
DDX21	gil1230564
H2A2	gil31979
HIST1H1C	gil4885375
HIST1H2BD	gil10800138
HNRNPAB	gil55956919
HNRNPC	gil306875
HNRNPD	gil119626277
HNRNPG	gil542850
HNRNPU	gil32358
HP1BP3	gil55961949
PBK1	gil3668141
RPL31	gil1655596
SEB4B	gil407421
SRSF10	gil5730079
SRSF7	gil72534660
TRA2B	gil4759098

(e)

Figure 3.1: hnRNPG preferentially binds an m^{6}A-modified hairpin in *MALAT1*. (a) Secondary structure of the 34-nucleotide hairpin derived from positions 2 505 – 2 538 of *MALAT1*, including the m^{6}A site at position 2 515 (red). The methylated form of the hairpin is termed 2 515-m^{6}A, and the unmethylated form is termed 2 515-A. (b) Gel shift showing binding of HeLa nuclear extract to the *MALAT1* hairpin in both its unmethylated (2 515-A) and methylated (2 515-m^{6}A) forms. (c) Left: Denaturing gel of the proteins pulled down by the unmethylated and methylated *MALAT1* hairpins. In the control, no RNA was used as bait. Right: Quantification of relative hnRNPG pull-down with the unmethylated and methylated hairpins, normalized to pulled-down Histone H1.2 (HIST1H1C). Data shown as mean; error bar = standard deviation; $n = 4$ biological replicates. (d) Quantitative mass spectrometry results showing proteins pulled down by the unmethylated (x-axis) and methylated (y-axis) *MALAT1* hairpins. hnRNPG (red) selectively bound to the methylated hairpin, while hnRNPU (blue) bound equally to both hairpins. (e) List of proteins pulled down by *MALAT1* hairpins and identified by mass spectrometry.

important role in neural development [247], and a frameshift mutation in the C-terminal region of hnRNPG causes an intellectual disability syndrome in humans [248].

3.2.2 hnRNPG binds the MALAT1 hairpin through a low-complexity region

hnRNPG is composed of an N-terminal globular RNA recognition motif of ∼90 amino acids, followed by ∼300 amino acids of low-complexity sequence, in which two thirds of the amino acid residues are serine, arginine, glycine, and proline (Figure 3.2(a)). In addition to the N-terminal RNA recognition motif (N-RRM), which binds A/C-rich sequences in single-stranded RNA [241], the C-terminal 58 amino acids of the low-complexity sequence have been shown to bind an RNA hairpin with an A/G-rich motif [249]. This C-terminal RNA binding domain (C-RBD) is highly conserved across species [249], and a frameshift mutation in this region causes an intellectual disability syndrome in humans [248]. However, the RNA sequence and structural preferences for the hnRNPG C-RBD have yet to be determined.

Due to the extensive low-complexity region of hnRNPG, the full-length recombinant protein was difficult to express and purify by us and others [249]. Instead, we purified each of the two known RNA binding domains of hnRNPG, the N-RRM and the C-RBD, fused to an N-terminal glutathione *S*-transferase. Similar to other low-complexity sequences [183, 224], the purified C-RBD self-assembled into higher-order structures, forming aggregates of ∼100 nm in size that were visible by electron microscopy (Figure 3.2(b)). We conducted gel shift assays to evaluate the binding of the N-RRM and C-RBD to the methylated and unmethylated *MALAT1* hairpins. While the N-RRM did not shift the *MALAT1* hairpin, the C-RBD shifted both hairpins with high cooperativity (Figure 3.2(c)). However, quantification was difficult to analyze due to the protein concentration dependent aggregation of the C-RBD. Therefore, we turned to ultraviolet cross-linking assays (Figure 3.2(d)). Similar to the gel shift assays, the N-RRM did not cross-link to the *MALAT1* hairpin, whereas the C-RBD cross-linked to both hairpins and more strongly to the m^6A-modified *MALAT1* hairpin.

The C-RBD contains three Arg-Gly-Gly (RGG) repeats. Since RGG motifs frequently

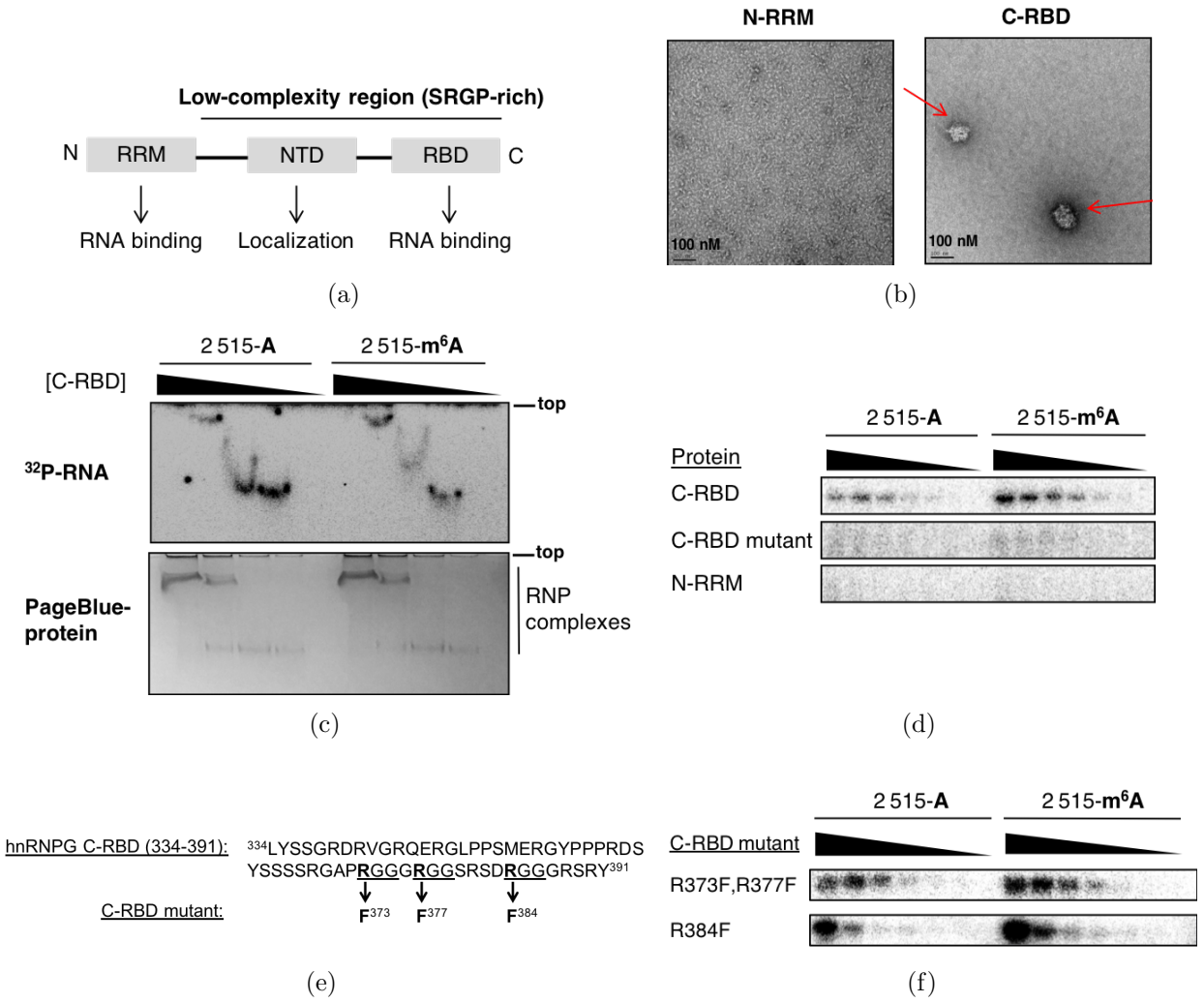


Figure 3.2: hnRNPG uses a low-complexity region to bind the *MALAT1* hairpin. (a) Domain structure of hnRNPG, including an N-terminal RNA recognition motif (RRM) and an SRGP-rich low-complexity region, which contains the nascent transcripts targeting domain (NTD) and a C-terminal RNA binding domain (RBD). (b) Electron microscopy images of the N-terminal RRM (N-RRM) and C-terminal RBD (C-RBD) of hnRNPG at 5 μM concentration. C-RBD aggregates are marked by red arrows. (c) Gel shift showing the ribonucleoprotein (RNP) complexes that form upon binding of the C-RBD of hnRNPG (0–20 μM) to the unmethylated and methylated *MALAT1* hairpins. The free RNA is not shown, as it has run much farther down the gel. Top: ^{32}P-labeled RNA gel; bottom: same gel stained for protein. (d) Ultraviolet cross-linking of the hnRNPG C-RBD, C-RBD mutant, and N-RRM (0–5 μM) to the unmethylated and methylated *MALAT1* hairpins. In the C-RBD mutant, all three RGG repeats in the C-RBD were mutated to FGG repeats. (e) hnRNPG C-RBD sequence with the three RGG motifs underlined and the introduced mutations shown. (f) Ultraviolet cross-linking of the hnRNPG C-RBD mutant R373F,R377F and mutant R384F (0–5 μM) to the unmethylated and methylated *MALAT1* hairpins.

function in RNA binding [206], we evaluated the role of the RGG motif by introducing arginine-to-phenylalanine mutations in the three RGG repeats of the C-RBD (Figure 3.2(d)–3.2(f)). Mutating one or two of the RGG motifs did not alter cross-linking efficiency. However, introducing mutations in all three RGG motifs abolished cross-linking to both the methylated and unmethylated *MALAT1* hairpins, suggesting that one or more of the RGG repeats in the C-RBD mediates *MALAT1* hairpin binding.

3.2.3 m^6A alters RNA structure and increases the accessibility of an hnRNPG binding motif

To examine transcriptome-wide binding sites of hnRNPG, we performed photoactivatable ribonucleoside-enhanced cross-linking and immunoprecipitation (PAR-CLIP) against hnRNPG in human embryonic kidney (HEK) 293T cells. hnRNPG PAR-CLIP of biological replicates yielded a total of 354 057 hnRNPG binding sites. AGRAC (R = A/G) was the most enriched motif among hnRNPG-bound RNAs, accounting for ∼30% (106 300) of hnRNPG binding sites identified by PAR-CLIP (Figure 3.3(a)). The AGRAC motif overlaps with the m^6A consensus motif RRACH, indicating that a large fraction of hnRNPG binding sites could be m^6A-modified.

The preferential binding of hnRNPG to m^6A-modified RNA can be mediated by direct recognition of m^6A, or through an effect of m^6A on the accessibility of its RNA binding motif. Mutating the m^6A site in the *MALAT1* hairpin to G, C, or U led to increased pull-down of hnRNPG from nuclear extract (Figure 3.3(b)). This result indicates that the preferential binding of hnRNPG to the methylated hairpin does not depend on the presence of the N^6-methyl group of the m^6A residue.

We then investigated whether m^6A modification influences the secondary structure of the *MALAT1* hairpin using structural mapping. We used S1 nuclease, which specifically cuts single-stranded regions, and RNase V1, which specifically cuts double-stranded, stacked regions of RNA (Figure 3.3(c)). Upon m^6A modification, the *MALAT1* hairpin became more

single-stranded in the region surrounding the m^{6}A site as shown by both increased S1 cuts and decreased V1 cuts in the region around the m^{6}A site. Our structural mapping results are consistent with m^{6}A disrupting the hairpin structure and increasing the accessibility of its surrounding nucleotides. The affected region in the *MALAT1* hairpin included the AGGAC

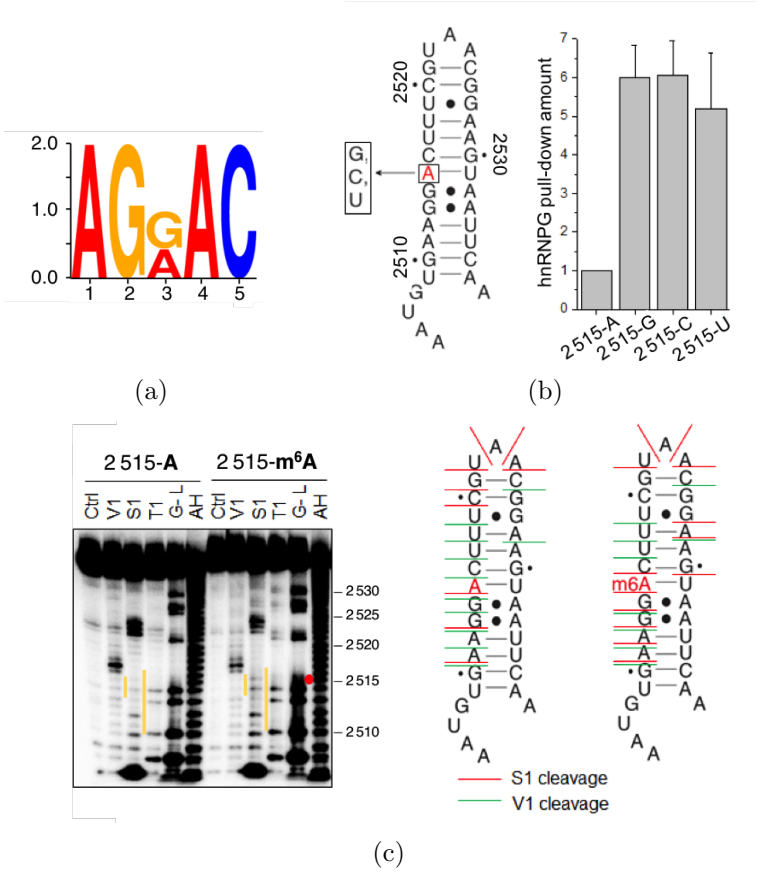


Figure 3.3: m^{6}A alters RNA structure to recruit hnRNPG. (a) Sequence logo of the most enriched motif within hnRNPG PAR-CLIP peaks. (b) Left: Secondary structure of the *MALAT1* hairpin, showing the A-2 515-to-G/C/U mutations that were introduced at the m^{6}A site. Right: Quantification of relative hnRNPG pull-down with the original (2 515-A) and mutated (2 515-G/C/U) *MALAT1* hairpins, normalized to pulled-down HIST1H1C. Data shown as mean; error bar = standard deviation; $n = 3$ biological replicates. (c) Left: Structural probing of the unmethylated and methylated *MALAT1* hairpins. The orange lines indicate regions with marked differences between the unmethylated and methylated hairpins. The location of the m^{6}A residue is indicated by a red dot. Ctrl, no nuclease added; V1; RNase V1 digestion; S1, S1 nuclease digestion; T1, RNase T1 digestion; G-L, G-ladder; AH, alkaline hydrolysis. Right: Secondary structure of the unmethylated and methylated *MALAT1* hairpins, marked at their S1 nuclease (red lines) and V1 nuclease (green lines) cleavage sites.

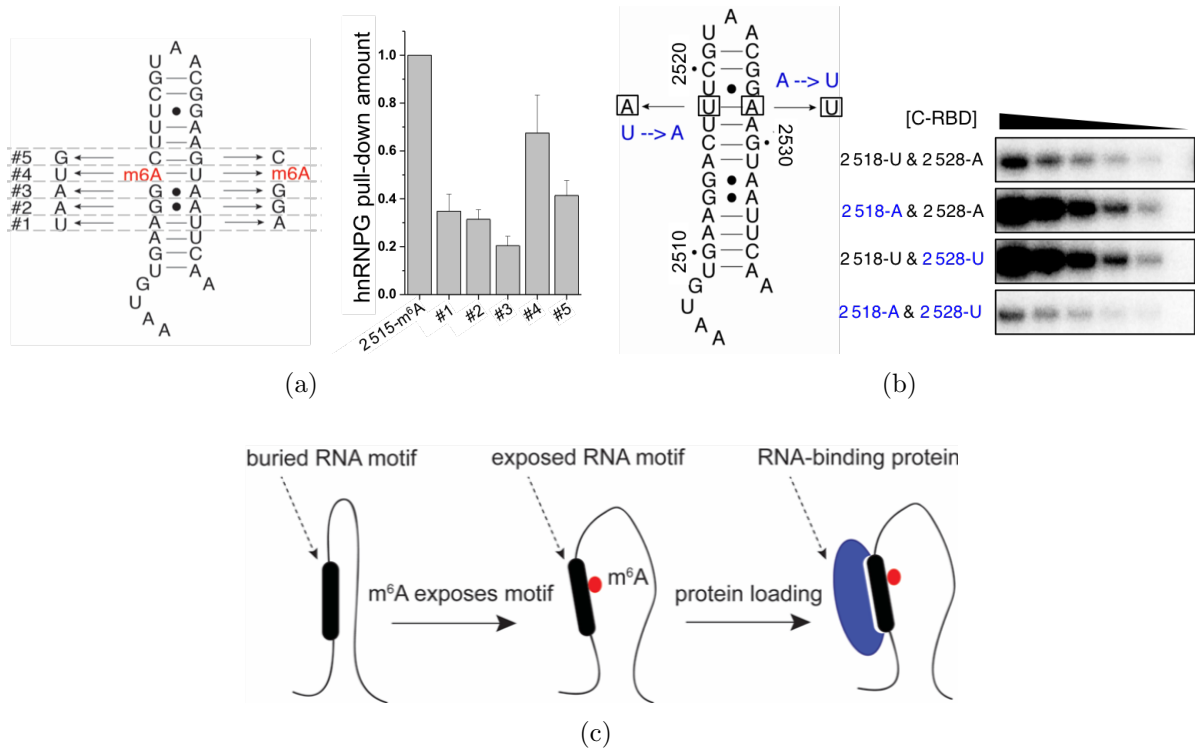


Figure 3.4: hnRNPG uses an m^{6}A-switch mechanism to bind the *MALAT1* hairpin. (a) Left: Secondary structure of the *MALAT1* hairpin, showing how each base pair involving the AGGAC sequence was switched individually in mutants #1–5. Right: Quantification of relative hnRNPG pull-down with the original (2 515-m^{6}A) and mutated (#1–5) *MALAT1* hairpins, normalized to HIST1H1C. Data shown as mean; error bar = standard deviation; $n = 2$ biological replicates. (b) Left: Secondary structure of the *MALAT1* hairpin, showing the U-2 515-to-A and A-2 528-to-U mutations that were introduced to destabilize hairpin structure. Right: Ultraviolet cross-linking of the hnRNPG C-RBD (0–2.5 μM) to the original (2 518-U and 2 528-A) and mutated (2 518-A and/or 2 528-U) *MALAT1* hairpins. (c) Model showing that m^{6}A disrupts RNA structure, exposes a motif that includes the m^{6}A site, and recruits an RNA-binding protein.

sequence, which matches the AGRAC motif found in the dominant hnRNPG binding sites identified by PAR-CLIP (Figure 3.3(a)).

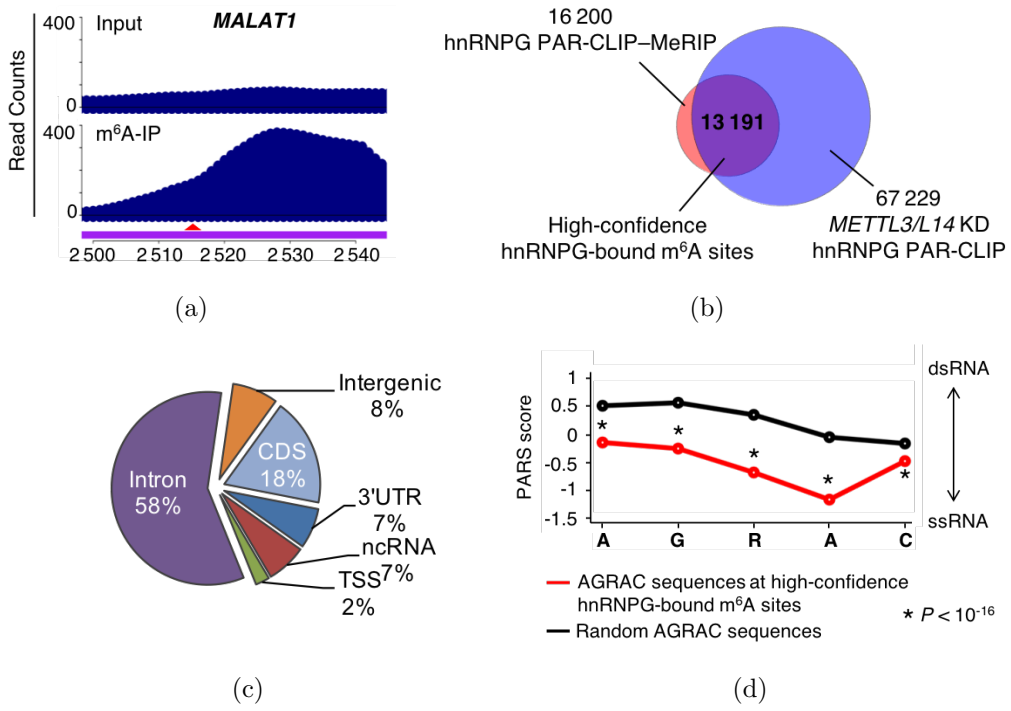
To evaluate the role of the AGGAC sequence for hnRNPG binding to the *MALAT1* hairpin, we switched individual base pairs involving the AGG[m^{6}A]C, in order to disrupt the sequence while minimizing the effect on hairpin structure. Each of these pairwise mutations in the *MALAT1* hairpin led to decreased hnRNPG pull-down from nuclear extract, with pairwise mutations at the m^{6}A position having the smallest effect (Figure 3.4(a)). This

result suggests that the AGG[m^{6}A]C sequence is important for hnRNPG binding, with the exception of the m^{6}A within this motif, which influences protein binding mainly by disrupting RNA structure to expose the surrounding bases. In addition, mutations that disrupted the structure of the *MALAT1* hairpin led to increased cross-linking of the purified C-RBD protein, while double mutations that restored the hairpin structure reduced cross-linking back to the wild-type hairpin level (Figure 3.4(b)). These results, together with the structural mapping data, indicate that m^{6}A modification of the *MALAT1* hairpin promotes hnRNPG binding by increasing the accessibility of the AGGAC motif (Figure 3.4(c)).

3.2.4 Transcriptome-wide identification of m^{6}A sites facilitating hnRNPG interactions

To identify m^{6}A-modified hnRNPG binding sites, we conducted hnRNPG PAR-CLIP followed by methylated RNA immunoprecipitation (MeRIP) of biological replicates [119]. In this hnRNPG PAR-CLIP–MeRIP experiment, hnRNPG-bound RNAs were isolated by PAR-CLIP ('input' sample), and then used as the input for MeRIP to enrich the m^{6}A-containing hnRNPG-bound RNAs ('IP' sample). hnRNPG PAR-CLIP–MeRIP yielded a total of 16 200 m^{6}A-modified hnRNPG binding sites ($log_2(IP/input) > 0.5$), including the m^{6}A at position 2 515 of *MALAT1* (Figure 3.5(a)). In addition, we conducted hnRNPG PAR-CLIP in cells depleted of either of the core m^{6}A methyltransferase components METTL3 and METTL14. Sites with decreased hnRNPG occupancy upon *METTL3* or *METTL14* knockdown were considered as m^{6}A methyltransferase-dependent hnRNPG binding sites. We identified 67 229 AGRAC motifs that were located within hnRNPG PAR-CLIP peaks but showed decreased hnRNPG binding following either *METTL3* or *METTL14* knockdown. Among these sites, 37 750 showed decreased hnRNPG binding in both *METTL3* and *METTL14* knockdown cells. 13 191 m^{6}A methyltransferase-dependent AGRAC sites were also identified as m^{6}A-modified hnRNPG binding sites by PAR-CLIP–MeRIP (Figure 3.5(b)–3.5(c)). We designate these 13 191 sites as high-confidence hnRNPG-bound m^{6}A sites.

Notably, using transcriptome-wide data from parallel analysis of RNA structure of human polyadenylated RNAs [250], we found that AGRAC sequences at high-confidence hnRNPG-bound m^{6}A sites were less structured than random AGRAC sequences (Figure 3.5(d)), suggesting that the m^{6}A-dependent structural change in the *MALAT1* hairpin could be generalized to hnRNPG-bound AGRAC motifs transcriptome-wide.



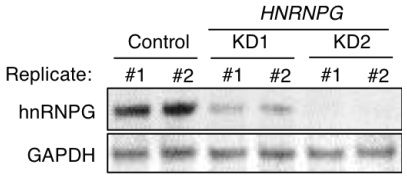
(a) (b)

(c) (d)

Figure 3.5: hnRNPG binds m^{6}A-modified RNAs transcriptome-wide. (a) PAR-CLIP–MeRIP input and IP (m^{6}A-IP) read counts in a region of the *MALAT1* transcript. The red arrowhead indicates the m^{6}A site at position 2 515. (b) Identification of high-confidence hnRNPG-bound m^{6}A sites (purple) as the overlap between m^{6}A-modified hnRNPG binding sites, identified by hnRNPG PAR-CLIP–MeRIP (pink), and m^{6}A methyltransferase-dependent hnRNPG-bound AGRAC sites, identified by hnRNPG PAR-CLIP in m^{6}A methyltransferase (*METTL3* and *METTL14*) knockdown HEK293T cells (blue). (c) Regional distribution of high-confidence hnRNPG-bound m^{6}A sites. (D) Comparison of the structure of AGRAC sequences at high-confidence hnRNPG-bound m^{6}A sites (red) vs. random AGRAC sequences (black) in human polyadenylated RNAs, based on parallel analysis of RNA structure (PARS) data [250]. The x-axis denotes nucleotide position; the y-axis shows the PARS score. Positive PARS scores indicate double-stranded conformation; negative scores indicate single-stranded conformation. P value, Mann-Whitney U test.

3.2.5 m^6A-dependent hnRNPG binding regulates gene expression and alternative splicing

hnRNPG regulates the transcription and alternative splicing of multiple genes [242, 243, 246]. To examine the effect of hnRNPG on gene expression and splicing transcriptome-wide, we conducted mRNA sequencing in HEK293T cells depleted of hnRNPG with two different siRNAs (Figure 3.6(a)). If hnRNPG binding to m^{6}A-modified RNAs contributes to its function as a regulator of transcription and splicing, then perturbations to cellular m^{6}A modifications should affect the expression and splicing of hnRNPG-regulated genes. To test this prediction, we compared our *HNRNPG* knockdown mRNA-seq data to our previous mRNA-seq data from cells with a global reduction in m^{6}A due to knockdown of *METTL3*

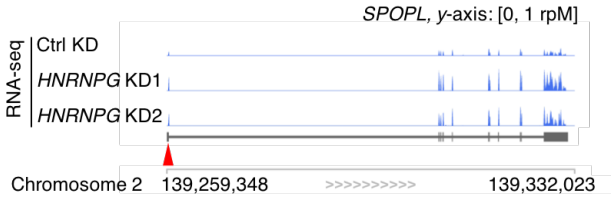


(a)

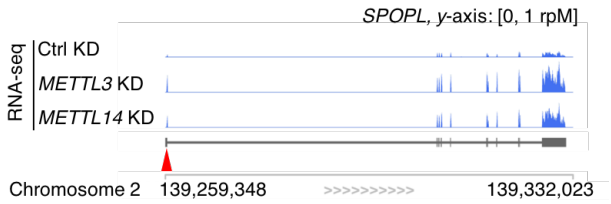
RNA abundance level changes		*METTL3/L14* KD / Control
HNRNPG **KD / Control**	All genes	336 + / 518 -
	Genes w/ HCG-m^{6}A	80 + / 169 -
HNRNPU **KD / Control**	All genes	4 + / 20 -
	Genes w/ HCG-m^{6}A	3 + / 5 -

+, co-up-regulation; -, co-down-regulation

(b)

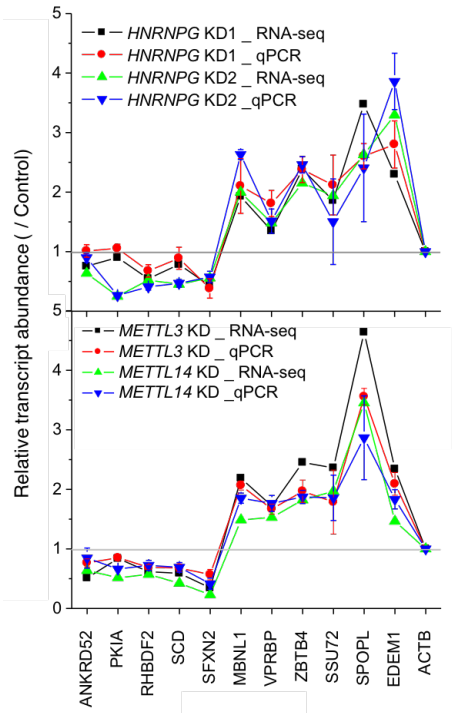


(c)



(d)

Figure 3.6: m^{6}A-dependent hnRNPG binding regulates mRNA abundance. (a) Western blot showing depletion of hnRNPG with two different siRNAs (KD1 and KD2) for mRNA-seq experiments. Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) was used as a loading control. (b) Number of genes with correlated changes in expression upon *HNRNPG* knockdown and m^{6}A methyltransferase (*METTL3* or *METTL14*) knockdown. HCG-m^{6}A, high-confidence hnRNPG-bound m^{6}A site. mRNA-seq data from *HNRNPU* knockdown HEK293T cells (Gene Expression Omnibus, GSE34995 [53]) were analyzed as a control. (c–d) mRNA-seq reads for *SPOPL* (speckle-type POZ protein-like) transcripts in control, *HNRNPG* knockdown (c), and m^{6}A methyltransferase knockdown (d) cells. The red arrowhead indicates the m^{6}A site.



(a)

mRNA abundance level changes	*METTL3/L14* KD/Ctrl
***HNRNPG* KD2 / Ctrl**	336+ / 518-
***HNRNPC* KD / Ctrl**	2725 + / 2526 -
***HNRNPG* KD2 / Ctrl and *HNRNPC* KD / Ctrl**	15 + / 24 -

"+" means co-up-regulate; "-" means co-down-regulate.

(b)

Figure 3.7: Validation of differential gene expression upon *HNRNPG* knockdown and m^{6}A methyltransferase knockdown. (a) Relative abundance of transcripts containing high-confidence hnRNPG-bound m^{6}A sites by RT–qPCR, in *HNRNPG* knockdown (top) and m^{6}A methyltransferase knockdown (bottom) HEK293T cells. Actin mRNA (ACTB) was used as control. (b) Number of genes for which changes in expression are correlated upon *HNRNPG*, *HNRNPC* (from reference [119]), and/or m^{6}A methyltransferase (*METTL3* and *METTL14*) knockdown.

or *METTL14* [119]. Using Cuffdiff2 [251], we found that hundreds of m^{6}A-containing RNA transcripts were similarly up- or down-regulated by both *HNRNPG* knockdown and m^{6}A methyltransferase knockdown (Figures 3.6(b)–3.6(d) and 3.7). As a control, gene expression changes in cells depleted of the mRNA binding protein hnRNPU did not correlate with gene expression changes upon *METTL3* or *METTL14* knockdown (Figure 3.6(b)), consistent with the observation that hnRNPU did not preferentially bind m^{6}A-modified RNA in nuclear pull-down assays (Figure 3.1(d)).

Next, we used DEXSeq [252] to examine changes in alternative splicing. Splicing changes resulting from *HNRNPG* knockdown were correlated with splicing changes resulting from m^{6}A methyltransferase knockdown (Figures 3.8(a) and 3.9(a)–3.9(b)). In particular, ∼1,000 exons in genes with high-confidence hnRNPG-bound m^{6}A sites were similarly up- or down-regulated by *HNRNPG* knockdown and m^{6}A methyltransferase knockdown (Figure 3.8(b)). For example, both *HNRNPG* knockdown and m^{6}A methyltransferase knockdown reduced the inclusion of an exon in the *NASP* transcript, which encodes nuclear autoantigenic sperm

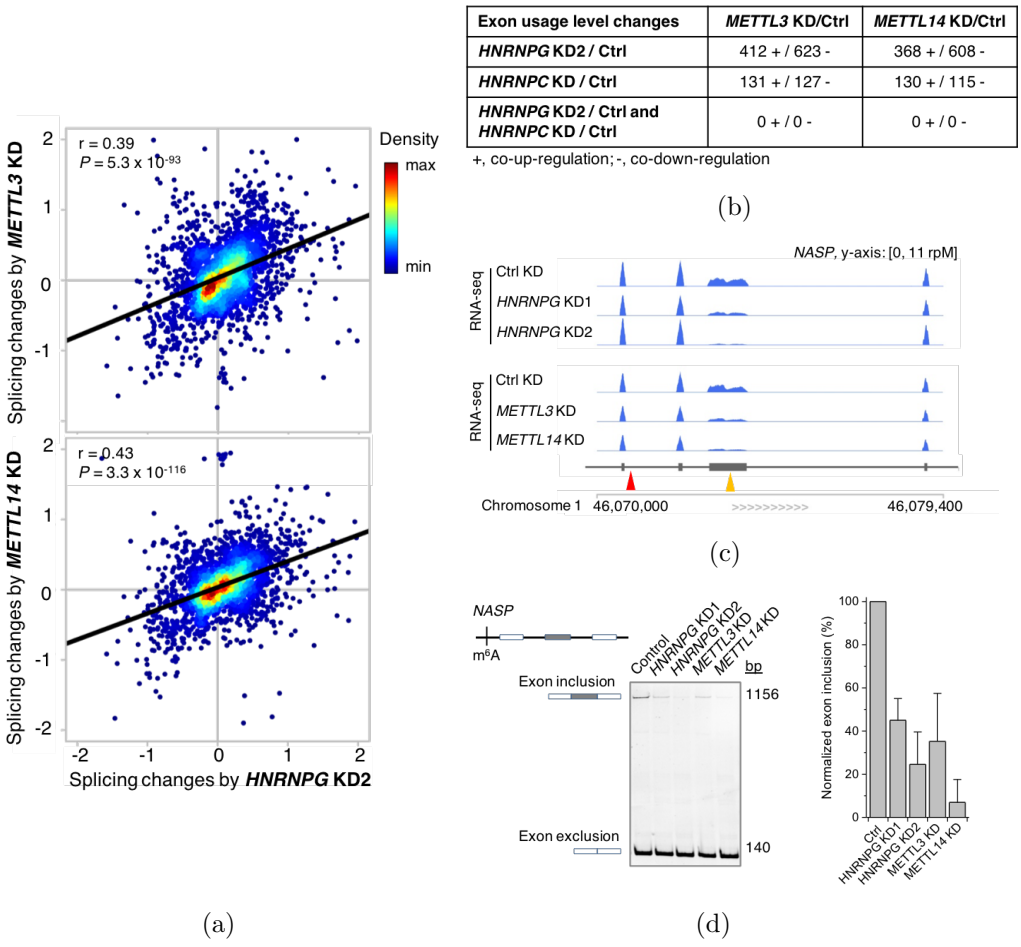


Figure 3.8: m^{6}A-dependent hnRNPG binding regulates alternative splicing. (a) Splicing changes of annotated differentially expressed exons upon *HNRNPG* knockdown with siRNA KD2 (x-axis) and m^{6}A methyltransferase (*METTL3* or *METTL14*) knockdown (y-axis), by log ratio of normalized counts relative to control knockdown, $log_2(KD/Control)$. Pearson's correlation coefficient r and P values are shown for each panel. (b) Number of exons for which changes in exon usage are correlated upon *HNRNPG*, *HNRNPC*, and/or m^{6}A methyltransferase (*METTL3* or *METTL14*) knockdown. For *HNRNPG* knockdown, only exons in genes with high-confidence hnRNPG-bound m^{6}A sites were counted. For *HNRNPC* knockdown, only exons in genes with high-confidence m^{6}A-switches were counted [119]. (C–D) mRNA-seq reads for *NASP* transcripts in control, *HNRNPG* knockdown (c) and m^{6}A methyltransferase knockdown (d) cells. The yellow arrowhead indicates the alternatively spliced exon; the red arrowhead indicates the m^{6}A site. (e) Reverse transcription and PCR (RT–PCR) validating differential exon usage in *NASP*. Data shown as mean; error bar = standard deviation; $n = 3$ biological replicates.

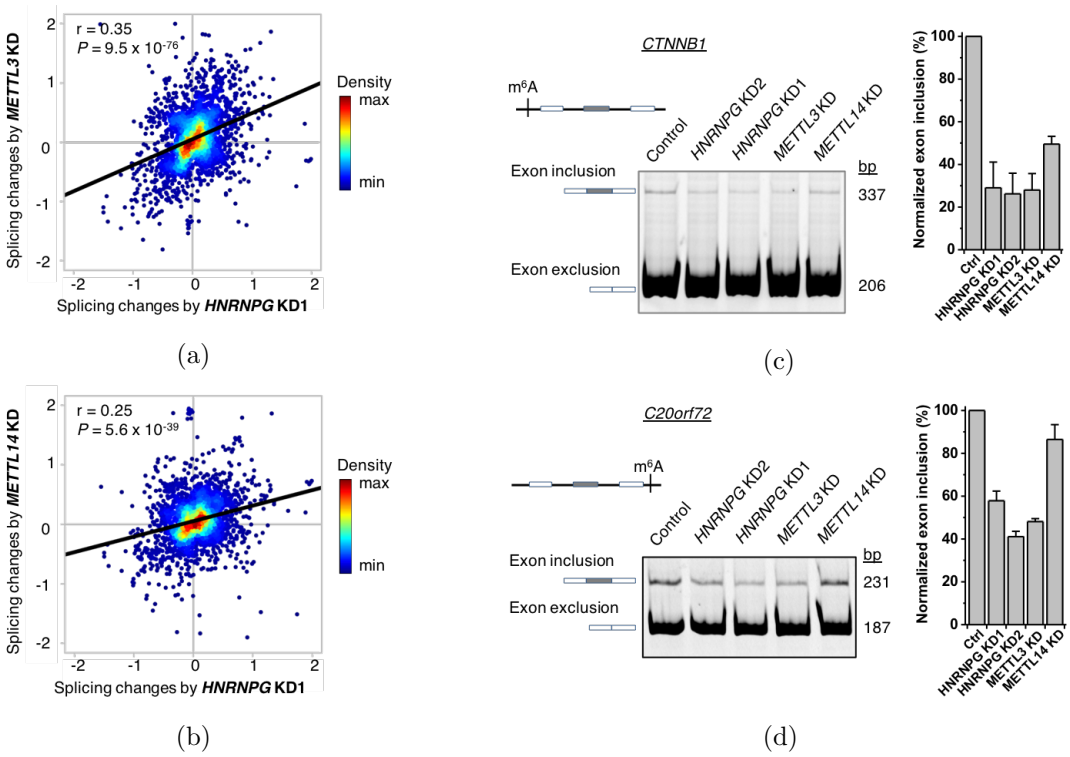


Figure 3.9: Validation of altered exon usage upon *HNRNPG* knockdown and m^{6}A methyltransferase knockdown. (a–b) Splicing changes of annotated differentially expressed exons upon *HNRNPG* knockdown with siRNA KD1 (x-axis) and m^{6}A methyltransferase (*METTL3* or *METTL14*) knockdown (y-axis), by $log_2(KD/Control)$. Pearson's correlation coefficient r and P values are shown for each panel. (C-D) RT–PCR validating differential exon usage in the *CTNNB1* (catenin beta 1) and *C20orf72* (mitochondrial genome maintenance exonuclease 1) transcripts upon *HNRNPG* knockdown and m^{6}A methyltransferase knockdown. Data shown as mean; error bar = standard deviation; $n = 3$ biological replicates.

protein (Figure 3.8(c)). hnRNPG PAR-CLIP–MeRIP revealed an m^{6}A-modified hnRNPG binding site in a nearby intron, suggesting that hnRNPG binding to this m^{6}A-modified site might promote inclusion of the alternatively spliced exon. The hnRNPG- and m^{6}A methyltransferase-dependent regulation of exon inclusion in *NASP* and two other transcripts was validated by RT–PCR (Figures 3.8(d) and 3.9(c)–3.9(d)). These results suggest that hnRNPG binding to m^{6}A-modified transcripts can regulate the alternative splicing of nearby exons.

The m^{6}A reader protein hnRNPC also binds m^{6}A-modified RNAs through an m^{6}A-switch mechanism. Previous results revealed thousands of RNA transcripts and hundreds of

exons that are similarly up- or down-regulated by *HNRNPC* knockdown and m^{6}A methyltransferase knockdown [119]. However, less than 10% of these transcripts and none of these exons were also co-regulated by *HNRNPG* and m^{6}A methyltransferase knockdown (Figures 3.7(b) and 3.8). Thus, while hnRNPC and hnRNPG both regulate gene expression and alternative splicing by binding to m^{6}A-modified transcripts, the target genes regulated by m^{6}A-dependent binding of hnRNPC are largely distinct from those regulated by m^{6}A-dependent binding of hnRNPG. These results suggest that m^{6}A-promoted hnRNPG and hnRNPC binding influence different subsets of gene transcripts.

3.3 Discussion

In this study, we identified hnRNPG as a new m^{6}A "reader" protein that uses a low-complexity region to bind m^{6}A-modified RNAs. Rather than directly recognizing the N^6-methyl group, hnRNPG binds a purine-rich motif that becomes accessible upon m^{6}A modification of an RNA hairpin. We identified 13 191 high-confidence hnRNPG-bound m^{6}A sites transcriptome-wide and showed that hnRNPG and m^{6}A modification cooperate to regulate gene expression and alternative splicing. These results demonstrate a role for the low-complexity domain of hnRNPG in the binding of m^{6}A-modified RNAs, and suggest that the m^{6}A-dependent binding of hnRNPG functions in the regulation of gene expression.

In the m^{6}A-switch mechanism, m^{6}A modulates protein binding by inducing an RNA structural change that alters the accessibility of a protein binding site. In the case of the m^{6}A reader hnRNPC, m^{6}A modification disrupts base pairing and exposes a U-rich binding site for hnRNPC in the opposite strand [1, 119]. hnRNPG binding to m^{6}A-modified RNA also depends on an m^{6}A-induced structural change. However, in contrast to hnRNPC, hnRNPG binds to a purine-rich motif that includes the m^{6}A site. Since the hnRNPG binding motif overlaps with the m^{6}A site, hnRNPG binding could compete with the binding of direct m^{6}A readers such as the YTH domain proteins. hnRNPG might also modulate the binding of other m^{6}A readers that use RNA structural changes as a readout for m^{6}A. Although

we found no indication that hnRNPG and hnRNPC cooperate to regulate the alternative splicing of m^{6}A-containing transcripts, positive or negative cooperativity between these and other m^{6}A readers may play a role in the regulation of other cellular processes.

Unlike previously described m^{6}A reader proteins, which all bind m^{6}A-modified RNA through folded domains, hnRNPG binds to m^{6}A-modified RNA through a low-complexity region that self-assembles into large particles *in vitro* (Figure 3.2(b)). Interactome studies have revealed that many RNA-binding proteins lack canonical RNA binding domains, and suggest that intrinsically disordered and low-complexity regions play an important role in RNA binding [193]. In fact, low-complexity sequences are prevalent in RNA-binding proteins, and an estimated one-third of RNA-binding proteins are highly disordered [192]. Unstructured RGG repeats are particularly common in RNA-binding proteins and have been shown to mediate structure- and sequence-specific RNA binding [206]. In addition to their widespread role in RNA binding, low-complexity domains frequently function in protein–protein interactions [193], are necessary and sufficient for the formation of granule-like structures *in vitro* [224], and are common sites of disease mutations that disrupt the balance between the assembly and clearance of ribonucleoprotein granules in the cell [193, 253]. Several other m^{6}A reader proteins contain low-complexity regions, but these regions have not been shown to directly bind m^{6}A-modified RNAs. The low-complexity regions of the YTH proteins function in their recruitment to various cellular bodies [118, 121, 123]. The RNA recognition motifs of the m^{6}A reader hnRNPA2B1 show only a slight preference for m^{6}A-modified RNAs *in vitro* [131], and it is yet to be determined whether this protein directly binds to m^{6}A. hnRNPA2B1 also has a glycine-rich low-complexity domain with four RGG motifs, which might contribute to the preferential binding of hnRNPA2B1 to m^{6}A-modified RNA. Moreover, hnRNPA2B1 may also use an m^{6}A-switch mechanism to selectively bind m^{6}A-modified RNAs. Thus, it is possible that hnRNPA2B1 and hnRNPG both use low-complexity regions and an m^{6}A-switch mechanism to bind m^{6}A-modified transcripts.

Our results reveal that hnRNPG regulates the alternative splicing of ∼1,000 exons in

m^{6}A-modified transcripts, but the mechanism for m^{6}A-dependent splicing regulation by hnRNPG remains unclear. hnRNPG might recruit or compete with splicing factors for binding to m^{6}A-modified transcripts, similar to the m^{6}A reader YTHDC1 [118]. In addition, both the N-terminal RRM and other parts of the low-complexity region of hnRNPG could contribute to the m^{6}A-dependent regulation of splicing. RNA-binding proteins frequently contain multiple repeats or combinations of RNA binding domains. This modular domain structure can contribute to RNA binding affinity or specificity, define the spacing between binding sites, organize RNA topology, or allow a single protein to simultaneously interact with multiple RNA molecules [191]. Given that the regulation of alternative splicing by RNA-binding proteins is highly position-dependent [43], the role of hnRNPG in splicing regulation is likely to be strongly dependent on the relationship between its different RNA binding domains.

The reversibility of m^{6}A modifications allows cells to dynamically regulate modification levels and fine-tune the fate of RNA transcripts. Indeed, modification fractions vary widely among different m^{6}A sites [72] and are regulated in response to external stimuli [60, 61]. The heterogeneity and dynamic nature of m^{6}A modification make it a useful mechanism for control of RNA fate. In addition to its role in binding m^{6}A-modified RNA, the low-complexity region of hnRNPG likely functions in granule formation. Cellular granules play an important role in regulating the localization of both RNA transcripts and RNA-binding proteins [253], which can in turn influence RNA splicing, translation, and turnover, all of which are processes regulated by m^{6}A [92, 93, 118, 121–123, 128–130]. The relationship between hnRNPG granule formation and binding to m^{6}A-modified RNAs could be a key to its function in the m^{6}A-dependent regulation of gene expression and alternative splicing.

Many cellular functions of m^{6}A are mediated by m^{6}A reader proteins. hnRNPG is the second m^{6}A reader protein shown to use the m^{6}A-switch mechanism [119], strongly suggesting that m^{6}A-mediated changes in RNA structure might impact interactions with many other RNA-binding proteins. At the same time, hnRNPG differs from previously

discovered m^6A reader proteins in that it uses a low-complexity region to bind to m^6A-modified RNA. Moreover, hnRNPG influences a large number of alternative splicing events in an m^6A-dependent manner. In the future, it will be important to investigate the mechanisms by which interactions between the low-complexity region of hnRNPG and m^6A-modified RNA modulate alternative splicing in the cell.

3.4 Materials and Methods

3.4.1 Mammalian cell culture, siRNA knockdown, and cell fractionation

Human cervical adenocarcinoma cell line HeLa (CCL-2) and human embryonic kidney cell line HEK293T (CRL-11268) were obtained from the American Type Culture Collection (ATCC) and cultured under standard conditions. Control siRNA (1027281, Qiagen), METTL3 siRNA (SI04317096, Qiagen), METTL14 siRNA (SI00459942, Qiagen), or hnRNPG siRNA (SI00700084 and SI00700077, Qiagen) was transfected into HEK293T cells at a concentration of 20-50 nM, using Lipofectamine RNAiMAX (13778100, Invitrogen) according to the manufacturer's instructions. Cells were collected 48 hours after transfection, shock-frozen in liquid nitrogen, and stored at -80 °C for further studies.

Nuclear and cytoplasmic extracts were isolated using the NE-PER Nuclear and Cytoplasmic Extraction Reagents (78833, Thermo Scientific) according to the manufacturer's instructions.

3.4.2 Western blotting

Western blots were performed using standard procedures. Briefly, 10-30 μg protein samples were separated on 4–12% polyacrylamide Bis-Tris gels (NP0336BOX, Invitrogen) and transferred to polyvinylidene fluoride membranes (IPVH00010, Millipore). The blots were probed with METTL3- (15073-1-AP, Proteintech), METTL14- (HPA038002, Sigma), hnRNPG- (sc-14581 and sc-48796, Santa Cruz Biotechnology), or GAPDH- (A00192-40, Genscript) specific

primary antibody, followed by rabbit anti-goat IgG-HRP (sc-2768, Santa Cruz Biotechnology) or goat anti-rabbit IgG-HRP (ab97051, Abcam) secondary antibody, and then visualized by enhanced chemoluminescence (RPN2109, GE Healthcare).

3.4.3 Protein expression

For expression of the N-terminal RNA recognition motif (N-RRM, residues 1–83) and C-terminal RNA binding domain (C-RBD, residues 334–391) of human hnRNPG protein, sequences encoding these hnRNPG domains were amplified by PCR from human HeLa cDNA libraries (637203, Clontech), and then subcloned into pGEX-6p-1 expression vectors using BamHI and XhoI restriction sites. Plasmid DNA was transformed into E. coli BL21-CodonPlus(DE3)-RP or BL21-CodonPlus(DE3)-RIL cells (Agilent). The transformed bacteria were grown to saturation at 37 °C, 200 rpm in Luria-Bertani Lennox medium with 100 μg/mL ampicillin, then diluted 1:100, grown in the same culture medium to an absorbance of $\sim$0.6 at 600 nm, and induced with 1 mM isopropyl β-D-1-thiogalactoside (IPTG). The bacteria were grown an additional 16–22 hours at 18 °C, 200 rpm, then harvested and sonicated at 4 °C. GST-fusion proteins were isolated from the soluble lysate using glutathione-Sepharose beads and stored in 10 mM Tris-Cl (pH 7.4), 100 mM KCl, 2.5 mM $MgCl_2$, 30% glycerol at −80 °C.

3.4.4 RNA oligos

The following RNA oligos were synthesized and purified by HPLC and/or denaturing gel electrophoresis, as previously described [239].

The RNA oligos used in Figures 3.1, 3.2, and 3.3(c):

2 515-A: 5′-AAUGUGAAGGACUUUCGUAACGGAAGUAAUUCAA-Biotin;

2 515-m^{6}A: 5′-AAUGUGAAGGm^{6}ACUUUCGUAACGGAAGUAAUUCAA-Biotin.

The RNA oligos used in Figure 3.3(b):

2 515-A: 5′-AAUGUGAAGGACUUUCGUAACGGAAGUAAUUCAA-Biotin;

2 515-G: 5′-AAUGUGAAGGGCUUUCGUAACGGAAGUAAUUCAA-Biotin;

2 515-C: 5′-AAUGUGAAGGCCUUUCGUAACGGAAGUAAUUCAA-Biotin;

2 515-U: 5′-AAUGUGAAGGUCUUUCGUAACGGAAGUAAUUCAA-Biotin.

The RNA oligos used in Figure 3.4(a):

#1: 5′-AAUGUGAUGGm^{6}ACUUUCGUAACGGAAGUAAAUCAA-Biotin;

#2: 5′-AAUGUGAAAGm^{6}ACUUUCGUAACGGAAGUAGUUCAA-Biotin;

#3: 5′-AAUGUGAAGAm^{6}ACUUUCGUAACGGAAGUGAUUCAA-Biotin;

#4: 5′-AAUGUGAAGGUCUUUCGUAACGGAAGm^{6}AAAUUCAA-Biotin;

#5: 5′-AAUGUGAAGGm^{6}AGUUUCGUAACGGAACUAAUUCAA-Biotin.

The RNA oligos used in Figure 3.4(b) were transcribed *in vitro* with T7 RNA polymerase and purified by denaturing gel electrophoresis:

2 518-U & 2 528-A: 5′-GAAAAAUGUGAAGGACUUUCGUAACGGAAGUAAUUCAAGATCA;

2 518-A & 2 528-A: 5′-GAAAAAUGUGAAGGACUAUCGUAACGGAAGUAAUUCAAGATCA;

2 518-U & 2 528-U: 5′-GAAAAAUGUGAAGGACUUUCGUAACGGUAGUAAUUCAAGATCA;

2 518-A & 2 528-U: 5′-GAAAAAUGUGAAGGACUAUCGUAACGGUAGUAAUUCAAGATCA.

3.4.5 RNA pull-down, gel shift, and cross-linking

The *in vitro* pull-down assays were performed as described [60]. The eluted protein samples were separated on 4–12% polyacrylamide Bis-Tris gels (NP0321BOX, Invitrogen) and stained with SYPRO Ruby Protein Gel Stain (S12000, Thermo Scientific) according to the manufacturer's instructions. Proteins in gel slices or in the entire pulled-down protein sample were digested with trypsin and identified using LC–MS/MS by the Donald Danforth Plant Science Center (Washington University, St. Louis, MO).

For the gel shift in Figure 3.1(b), the gel-purified 5′ ^{32}P-labeled RNA oligos were refolded by heating at 90 °C for 1 min, then at 30 °C for 5 min. 3 μL nuclear extract and 6 μL refolded RNA were incubated together at room temperature for 30 min, and then at 4 °C for 2 hrs. Each sample was mixed with 1 μL 50% glycerol, separated on an 8% polyacrylamide, 44.5

mM Tris-borate (pH 8.3 at 25 °C, pH 8.9 at 5 °C), and 1 mM Na_2EDTA (ethylenediaminetetraacetic acid) native gel at 5 °C, and visualized by phosphorimaging using the Personal Molecular Imager (Bio-Rad).

For the gel shift and ultraviolet cross-linking assays in Figures 3.2 and 3.4(b), gel-purified 5′ ^{32}P-labeled RNA oligos were refolded by incubating at 90 °C for 1 min, then at room temperature for 3 min. The refolded RNA was combined with purified recombinant N-RRM, C-RBD, or C-RBD mutant protein either in 10 mM Tris-Cl (pH 7.4), 100 mM KCl, 2.5 mM $MgCl_2$ (gel shift) or in 10 mM HEPES (pH 8.0), 50 mM KCl, 1 mM EDTA, 0.05% Triton X-100, 5% glycerol, 1 mM DTT, 10 μg/mL salmon sperm DNA (cross-linking), and the RNA–protein mixture was incubated for 30 min at 30 °C, 600 rpm. For the gel shift assay in Figure 3.2(c), each sample was mixed with 1 μL 50% glycerol, separated on an 8% polyacrylamide, 44.5 mM Tris-borate (pH 8.3 at 25 °C, pH 8.9 at 5 °C), and 1 mM Na_2EDTA native gel at 5 °C, and visualized by phosphorimaging using the Personal Molecular Imager (Bio-Rad). Ultraviolet cross-linking was conducted at 254 nm, 150 mJ/cm^2. The cross-linked protein samples were separated on 4–12% polyacrylamide Bis-Tris gels, visualized by phosphorimaging, and stained in PageBlue Protein Staining Solution (24620, Thermo Scientific).

3.4.6 RNA structural probing by RNase V1/S1

The synthetic RNA oligos were 5′ end-labeled with γ^{32}P-ATP by T4 PNK (70031, Affymetrix), gel purified, and re-folded. Structural probing assays with RNase T1, nuclease S1 and RNase V1 were performed as previously described [72].

3.4.7 Electron microscopy

The purified GST-fusion C-RBD or N-RRM domain of the hnRNPG protein was diluted to 5 μM in 10 mM Tris-Cl (pH 7.4), 100 mM KCl, 2.5 mM $MgCl_2$. The 5 μL sample solution was spotted on a 400 mesh carbon grid, washed with 1–2 drops of water, stained with 2 drops of

1% uranyl acetate, blotted, and dried in air. Images were obtained at 68.6 K× magnification on an FEI Tecnai G2 F30 Super Twin scanning transmission electron microscope.

3.4.8 PAR-CLIP and PAR-CLIP–MeRIP

PAR-CLIP was performed as previously reported [254] with the following modifications. HEK293T cells in 15-cm plates treated following normal PAR-CLIP procedures were lysed and digested with a combination of RNase I (AM2295, Ambion, 12 μL diluted 1:50 with H_2O) and Turbo DNases (2 μL) for 3 min at 37 °C, shaking at 1 100 rpm. The lysate was then immediately cleared by spinning at 14 000 rpm, 4 °C for 30 min, and placed on ice for further use. hnRNPG binding sites were identified by PARalyzer v1.1 with default settings [255].

The PAR-CLIP–MeRIP experiment applied m^6A-antibody immunoprecipitation [256] to the hnRNPG PAR-CLIP RNA samples (from HEK293T cells in eight 15-cm plates). The hnRNPG PAR-CLIP RNA sample was incubated with m^6A-specific antibody (202003, SYSY), RNase inhibitor (80 units, Sigma-Aldrich), human placental RNase inhibitor (NEB) in 200 μL 1× IP buffer (50 mM Tris-Cl (pH 7.4), 750 mM NaCl and 0.5% (vol/vol) Igepal CA-630) at 4 °C for 2 hours under gentle shaking conditions. For each PAR-CLIP–MeRIP experiment, 20 μL Protein A beads (10002D, Thermo Scientific) were washed twice with 1 mL 1× IP buffer, blocked by a 2-hour incubation with 100 μL 1× IP buffer supplemented with BSA (0.5 mg/ml), RNasin, and Human placental RNase inhibitor, and then washed twice with 100 μL 1× IP buffer. The pre-blocked Protein A beads were then combined with the prepared immuno-reaction mixture and incubated at 4 °C for 2 hours, followed by three washes with 100 μL 1× IP buffer. Finally, the RNA was eluted by 1-hour incubation with 20 μL elution buffer (1× IP buffer and 6.7 mM m^6A, Sigma-Aldrich) under gentle shaking conditions, and then purified by ethanol precipitation. The purified RNA sample (IP) as well as the input PAR-CLIP RNA sample (Input control) were used for library construction.

Libraries for both PAR-CLIP and PAR-CLIP–MeRIP experiments were prepared using

the TruSeq Small RNA Sample Preparation Kit (RS-200-0012, Illumina) according to the manufacturer's instructions, and then sequenced by Illumina HiSeq 2000 with single-end 50-bp reads. The control and IP samples from PAR-CLIP–MeRIP experiments were sequenced together in two lanes of one flow cell, and the reads from two lanes of each sample were combined for analysis. (The control and knockdown (KD) samples from *METTL3/L14* KD experiments were sequenced and combined for analysis in the same manner.) The raw sequencing data was trimmed using the Trimmomatic computer program version 0.30 to remove adaptor sequences, and mapped to the human genome version hg19 by Bowtie 1.0.0 [257] without any gaps and allowing for at most two mismatches. Approximately 20 million reads were mapped to hg19 for each sample.

3.4.9 Detection and distribution analysis of m^{6}A sites within hnRNPG binding sites

Detection of hnRNPG-bound m^{6}A sites by PAR-CLIP–MeRIP involves comparing the read counts of the IP sample with those of the control (Ctrl) sample as follows: (1) we identified all AGRAC motifs within hnRNPG PAR-CLIP peaks; (2) we performed transcriptome-wide scanning to compare read counts of each AGRAC motif in (1) from both Control and IP samples to calculate the fold change score, score = $log_2(Counts_{\mathrm{IP}}/Counts_{\mathrm{Control}})$. AGRAC motifs with $log_2(Counts_{\mathrm{IP}}/Counts_{\mathrm{Control}})$ larger than 0.5 were considered to be hnRNPG-bound m^{6}A sites.

Detection of hnRNPG-bound m^{6}A sites by PAR-CLIP involves comparing the read counts of the *METTL3/L14* KD sample with that of the control (Ctrl) sample as follows: (1) we identified all AGRAC motifs within hnRNPG PAR-CLIP peaks; (2) we performed transcriptome-wide scanning to compare read counts of each AGRAC motif in (1) from both Control and *METTL3/L14* KD samples to calculate the fold change score, score = $log_2(Counts_{\mathrm{KD}}/Counts_{\mathrm{Control}})$. AGRAC motifs with $log_2(Counts_{\mathrm{KD}}/Counts_{\mathrm{Control}}) < -0.5$ were considered to be hnRNPG-bound m^{6}A sites.

High-confidence m^{6}A (HC m^{6}A) sites within hnRNPG binding sites fulfill the following two requirements: (1) $log_2(Counts_{\mathrm{IP}}/Counts_{\mathrm{Control}}) > 0.5$; (2) $log_2(Counts_{\mathrm{KD}}/Counts_{\mathrm{Control}}) < -0.5$. Pie charts illustrating the distribution of high-confidence hnRNPG-bound m^{6}A sites within each segment were made using the following hierarchy: intron > ncRNA > 3 UTR > 5 UTR > CDS > intergenic.

3.4.10 RNA sequencing, graphics, and statistical analysis

RNA-seq experiments were performed on two replicates of RNA samples from *HNRNPG* knockdown (KD1 and KD2) as well as control HEK293T cells 48 hours after transfection. Total RNA was extracted using the RNeasy Plus kit (74104, Qiagen). Libraries were prepared using the TruSeq Stranded mRNA LT Sample Prep Kit (RS-122-9005DOC, Illumina). KD and control samples were sequenced together in four lanes in one flow cell. All samples were sequenced by Illumina HiSeq 2000 with paired-end 100-bp reads. The reads from the four lanes of each sample were combined for all analyses. The RNA-seq data was mapped using the splice-aware alignment algorithm TopHat version 1.1.4 [258] based on the following parameters: tophat –num-threads 8 –mate-inner-dist 200 –solexa-quals –min-isoform-fraction 0 –coverage-search-segment-mismatches 1. Gene expression level changes were analyzed using Cuffdiff2 [251]. Approximately 140–200 million reads were mapped for each sample. Differential splicing was determined using DEXSeq [252] based on Cufflinks-predicted, non-overlapping exons.

Sequence logos were generated using the WebLogo package. R statistical package was used for all statistical analysis (unless stated otherwise).

3.4.11 RT–PCR quantification

Total RNA was extracted from HEK293T cells and reverse transcribed using the SuperScript III First-Strand Synthesis System (18080-051, Life Technologies). In order to validate the splicing changes identified from our RNA-seq data, we performed RT–PCR measurements

using Taq DNA Polymerase (Thermo Scientific) under the following conditions: 95 °C for 3 min; 30 cycles of [95 °C for 30 s, 55 °C for 30 s, 72 °C for 1 min]; 72 °C for 10 min. For the target alternatively spliced exon, we designed and used primers annealing to both neighboring constitutive exons. The PCR products were separated on a 10% polyacrylamide, 89 mM Tris-borate (pH 8.3 at 25 °C), and 2 mM Na_2EDTA gel and stained with SYBR Gold Nucleic Acid Gel Stain (S11494, Thermo Scientific).

To validate the gene expression level changes identified from our RNA-seq data, we performed RT–qPCR measurements using Power SYBR Green PCR Master Mix (4367659, Thermo Scientific) under the following conditions: 50 °C for 3 min; 95 °C for 10 min; 40 cycles of [95 °C for 15 s, 60 °C for 1 min]; 40 °C for 1 min; 95 °C for 15 s; 60 °C for 30 s.

The primer sequences are listed below (*Gene name*: forward primer; reverse primer):

C20orf72: ACAGCGGATGATTCTGGAAC; TTCCTGGGGTGAAAGTATGC

NASP: TGTGCATGTGGAAGAGGAAG; GAAGGTGTGCATGTGGAAGA

CTNNB1: GAAAATCCAGCGTGGACAAT; CAGGACTTGGGAGGTATCCA

PKIA (1): CCTGGTTTCCCCAAAGAAGT; TGATTGGAAACCTTCTTGTCTTT

PKIA (2): TGGTAGCAATGACTGATGTGG; ACTTGCAGAGGAAACCAGGA

RHBDF2 (1): AGAGCCAGAGACCCAAGACA; CCCAAGACTCAGAGAGGCA

RHBDF2 (2): GAGTACCCAGGAAGCTGCAC; TACAGATGCTCCGGTGTCAA

SCD (1): TGTTCGTTGCCACTTTCTTG; GGGGGCTAATGTTCTTGTCA

SCD (2): CTCCACTGCTGGACATGAGA; AATGAGTGAAGGGGCACAAC

SFXN2 (1): GCCAGACTGGTCTCGAACTC; ACGGTCCCCTTTTTAGCACT

SFXN2 (2): CCTGGGATTGGTCGAAAAG; AAATGCCACCAGTTACAGCC

ANKRD52 (1): CTGTGCCGAGACTTTAAGGG; GCGAGTATCCGCTGTAATCC

ANKRD52 (2): AGACGCTGGTGAATCTGGAC; GCTGTAAGCACCTCCACACA

MBNL1 (1): AATATCTTCATCCACCCCCA; TTGGCTAGTTGCATTTGCTG

MBNL1 (2): GCTGCATCTGTCTATGCCAA; CGAATTTCCAAGCTGCTTTC

VPRBP (1): GCTGACAAAAGAGGCTGACC; GCTGAGGATGAGCAGTAGGG

VPRBP (2): TGATAGAATATGGCCCAGCG; CCAATTGCAGGCAATAGAAA

ZBTB4 (1): TTCCATGCCTTTGGATCTTC; ATTTGGGGGTCAAGATAGGG

ZBTB4 (2): GCTCACTTCAGCCCCACTAC; AGACGAGGAAGAGGAGGAGG

SSU72 (1): GCACTTCCCGACATACCTGT; GCACAATGACAGCAGCATCT

SSU72 (2): AAATAAGAGAATCAAGCCCCG; TTCCACCACCTGGTCATACA

SPOPL (1): GCTGGAGTCGTAACTCGGAA; CGCTCCTAAACTTCTTCCCC

SPOPL (2): GGAGGTTTGTTCTGGTGCAT; GCCCTTAAGAAGCACACTGG

EDEM1 (1): AGCCTCCTTTCTGCTCACAG; GGTGTTTTCAAAAGCAGGGA

EDEM1 (2): ATGAGCATCTTCGGGAATTG; AACTCATGAGGTTTCGGCCT

3.4.12 Data Deposition

The sequencing data have been deposited to National Center for Biotechnology Information Gene Expression Omnibus database under accession number GSE74085.

Chapter 4

Co-transcriptional m^{6}A-dependent Gene Regulation by hnRNPG

Acknowledgement: This chapter is derived from a manuscript in preparation. The authors of that manuscript are: Katherine I. Zhou, Żaneta Matuszek, Jessica N. Pan, Marc Parisien, and Tao Pan. Author contributions: Conceptualization, K.I.Z. and T.P.; Methodology, K.I.Z. and T.P.; Software, M.P.; Formal Analysis, K.I.Z. and M.P.; Investigation, K.I.Z. (immunoprecipitation, immunofluorescence, pull-down, limited proteolysis, SPR, DLS, and sequencing experiments), Ż.M. (immunoprecipitation experiments), and J.N.P. (cloning); Writing – Original Draft, K.I.Z.; Writing – Review & Editing, K.I.Z., M.P., and T.P.; Supervision, T.P.

4.1 Introduction

Messenger RNAs (mRNAs) undergo both co-transcriptional and post-transcriptional processing. In eukaryotes, the carboxy-terminal domain (CTD) of RNA polymerase II (RNAPII) couples mRNA processing to transcription by recruiting constitutive mRNA processing factors [31]. While the CTD also contributes to alternative splicing, few alternative splicing factors have been shown to directly interact with the CTD [12, 54]. Moreover, local features of chromatin or of nascent mRNA likely act together with the CTD to target alternative splicing factors to specific splicing events. In particular, the mRNA modification m^{6}A occurs co-transcriptionally [85, 86] and can influence alternative splicing [149], so m^{6}A might co-transcriptionally regulate alternative splicing factor binding to nascent transcripts. Several RNA-binding proteins that preferentially bind to m^{6}A-modified transcripts, called m^{6}A reader proteins, influence alternative splicing in an m^{6}A-dependent manner [3, 118, 119]. Knockdown of the m^{6}A reader protein heterogeneous nuclear ribonucleoprotein G (hnRNPG) leads to splicing changes that correlate with splicing changes upon m^{6}A methyltransferase knockdown [3]. However, the mechanisms for m^{6}A-dependent regulation of alternative splic-

ing by hnRNPG remain unknown. hnRNPG is unique among m^6A reader proteins in that it uses a low-complexity region to selectively bind m^6A-modified RNAs [3]. Here, we investigate how the CTD of RNAPII, the low-complexity region of hnRNPG, and m^6A modification of nascent mRNA act together to regulate alternative splicing.

4.2 Results

4.2.1 hnRNPG interacts with RNA polymerase II in vivo

We found that hnRNPG fractionated with chromatin, and that RNAPII co-immunoprecipitated with hnRNPG from whole cell or chromatin extracts from human embryonic kidney (HEK) 293T cells (Figure 4.1(a)–4.1(d)). Since the chromatin extracts were treated with micrococcal nuclease, this result also demonstrated that the interaction between hnRNPG and RNAPII did not depend on exposed DNA or RNA. Upon transcription inhibition with actinomycin D (actD) or 5,6-dichloro-1-β-D-ribofuranosylbenzimidazole (DRB), co-immunoprecipitation of RNAPII with hnRNPG decreased (Figure 4.1(e)), whereas upon inhibition with camptothecin (CPT), co-immunoprecipitation of RNAPII increased (Figure 4.1(f)). While all three inhibitors block transcription elongation, actinomycin D and DRB decrease occupancy of transcribing RNAPII on chromatin, whereas camptothecin leads to the accumulation of stalled RNAPII [259, 260]. Therefore, hnRNPG likely interacts with transcribing, chromatin-bound RNAPII. Consistent with this result, inhibiting RNAPII transcription altered the localization of hnRNPG. While hnRNPG in untreated cells localized to small nucleoplasmic granules, treatment with α-amanitin or actinomycin D led to the re-localization of hnRNPG to dense clusters in the nucleus (Figure 4.2(a)).

Next, we examined the function of the different regions of hnRNPG in the interaction with RNAPII. The hnRNPG protein is composed of an ∼80-amino-acid RNA recognition motif (RRM) and a ∼300-amino-acid low-complexity sequence, which includes two regions (RGG1 and RGG2) that each contain three Arg-Gly-Gly (RGG) repeats. RGG repeats

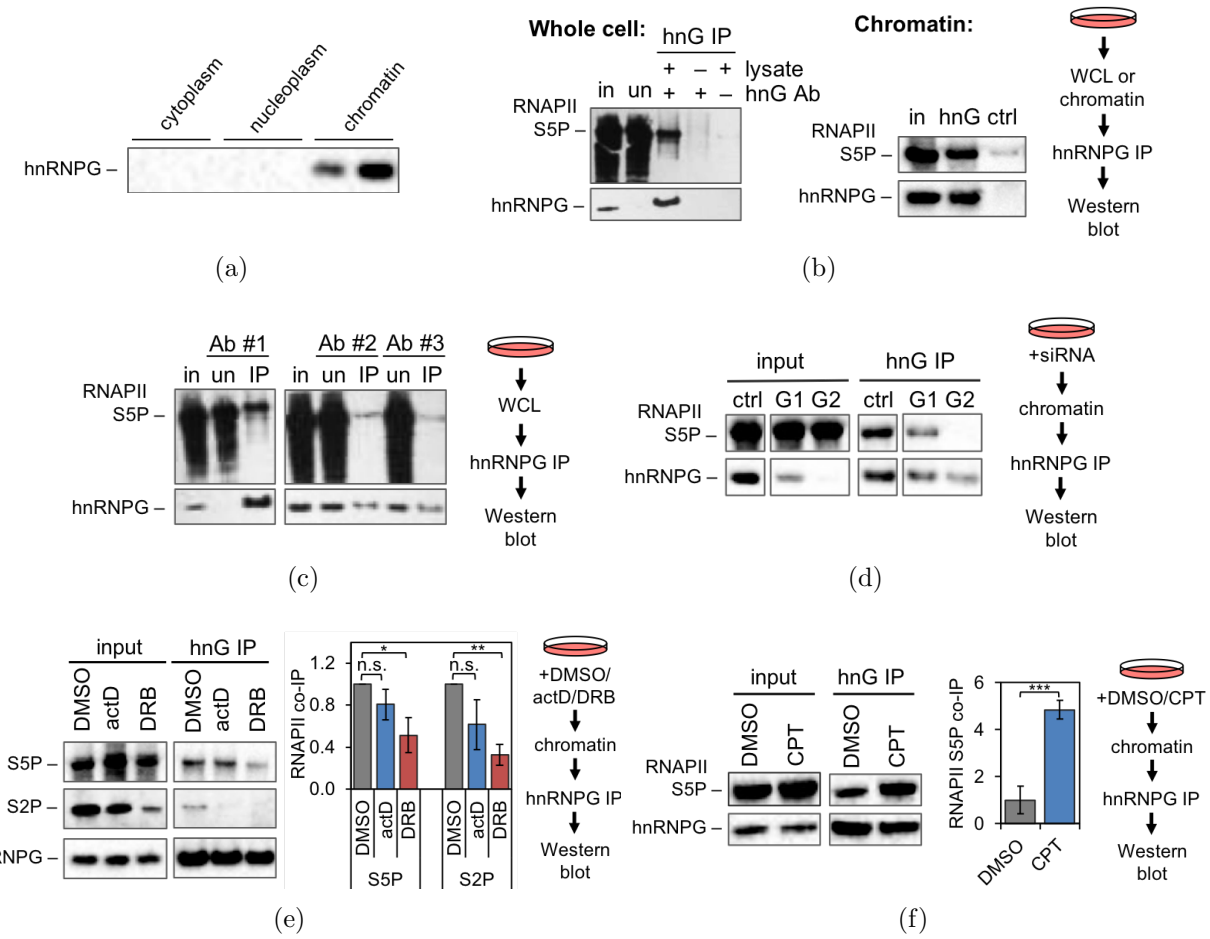


Figure 4.1: Interaction of hnRNPG with RNA polymerase II in cells. (a) Western blot showing hnRNPG in the chromatin fraction, but not in the cytoplasmic or nucleoplasmic fraction, of HEK293T cells. (b) Western blot showing co-immunoprecipitation (co-IP) of Ser5-phosphorylated RNAPII (RNAPII S5P) with hnRNPG in HEK293T whole cell or chromatin extracts. in: input (whole cell or chromatin extract); un: unbound fraction; hnG IP: hnRNPG IP; hnG Ab: hnRNPG antibody; hnG: hnRNPG IP; ctrl: isotype control IP. (c) Western blot showing co-IP of RNAPII S5P in whole cell lysates of HEK293T cells upon immunoprecipitation of hnRNPG using three different antibodies (Ab #1–3). in: input (whole cell lysate); un: unbound; IP: hnRNPG IP. (d) Western blot showing co-IP of RNAPII S5P with hnRNPG in chromatin extracts of HEK293T cells transfected with control siRNA (ctrl) or one of two different hnRNPG siRNAs (G1 and G2). (e) Western blot showing co-IP of RNAPII with hnRNPG in chromatin extracts of HEK293T cells treated with 0.5% v/v DMSO, 5 μg/mL actinomycin D (actD), or 100 μM DRB for 2 hours. (f) Western blot showing co-IP of RNAPII S5P with hnRNPG in chromatin extracts of HEK293T cells treated with 0.12% v/v DMSO or 6 μM camptothecin (CPT) for 5 hours. For d–f: input, chromatin extract; hnG IP, hnRNPG IP. Bar graphs: RNAPII co-IP was measured relative to RNAPII co-IP for DMSO-treated cells. Error bars, $\pm$ 1 standard deviation; $n = 3$ biological replicates; n.s., not significant; * $p < 0.05$, ** $p < 0.01$; *** $p < 0.001$ by two-tailed t-test.

are commonly found in RNA-binding proteins and frequently function in protein–protein and RNA–protein interactions [206], and the RGG2 region in particular can mediate the preferential binding of hnRNPG to m^6A-modified RNA [3]. We introduced several point mutations in either the RRM, RGG1, or RGG2 region of hnRNPG to generate the mutants RRMmut, RGG1mut, and RGG2mut, respectively (Figure 4.2(b)). We knocked down endogenous hnRNPG and overexpressed FLAG-tagged forms of either wild-type or mutant hnRNPG (FLAG–hnRNPG) in HEK293T cells. Wild-type and mutant FLAG–hnRNPG had similar stability since they were expressed at similar levels (Figure 4.2(c)). RNAPII co-immunoprecipitated with wild-type FLAG–hnRNPG, whereas point mutations in the RRM, RGG1, or RGG2 region decreased co-immunoprecipitation (Figure 4.2(c)). The mutations in RRMmut have previously been shown to abolish RNA binding by the RRM region of hnRNPG [261], raising the possibility that the RRM region increases the association of hnRNPG with RNAPII by enhancing the recruitment of hnRNPG to nascent RNA at transcription sites. Since hnRNPG is an m^6A reader protein, m^6A modifications in nascent RNA could also influence the recruitment of hnRNPG to transcribing RNAPII. In fact, camptothecin enhances the interaction of METTL3 with RNAPII, possibly leading to increased m^6A modification of nascent transcripts [89], so the enhanced interaction between hnRNPG and RNAPII upon camptothecin treatment (Figure 4.1(f)) could be partly due to increased m^6A modification of nascent RNA. To test this possibility, we measured the interaction between hnRNPG and RNAPII upon knockdown of the m^6A methyltransferase catalytic subunit, methyltransferase-like 3 (METTL3), which associated with RNAPII even in the absence of camptothecin (Figure 4.2(d)). METTL3 knockdown slightly decreased the association of hnRNPG with RNAPII, but this effect was not significant (Figure 4.2(e)). Therefore, the effect of camptothecin on m^6A modification is unlikely to fully explain the large effect of camptothecin treatment on the interaction between hnRNPG and RNAPII (Figure 4.1(f)).

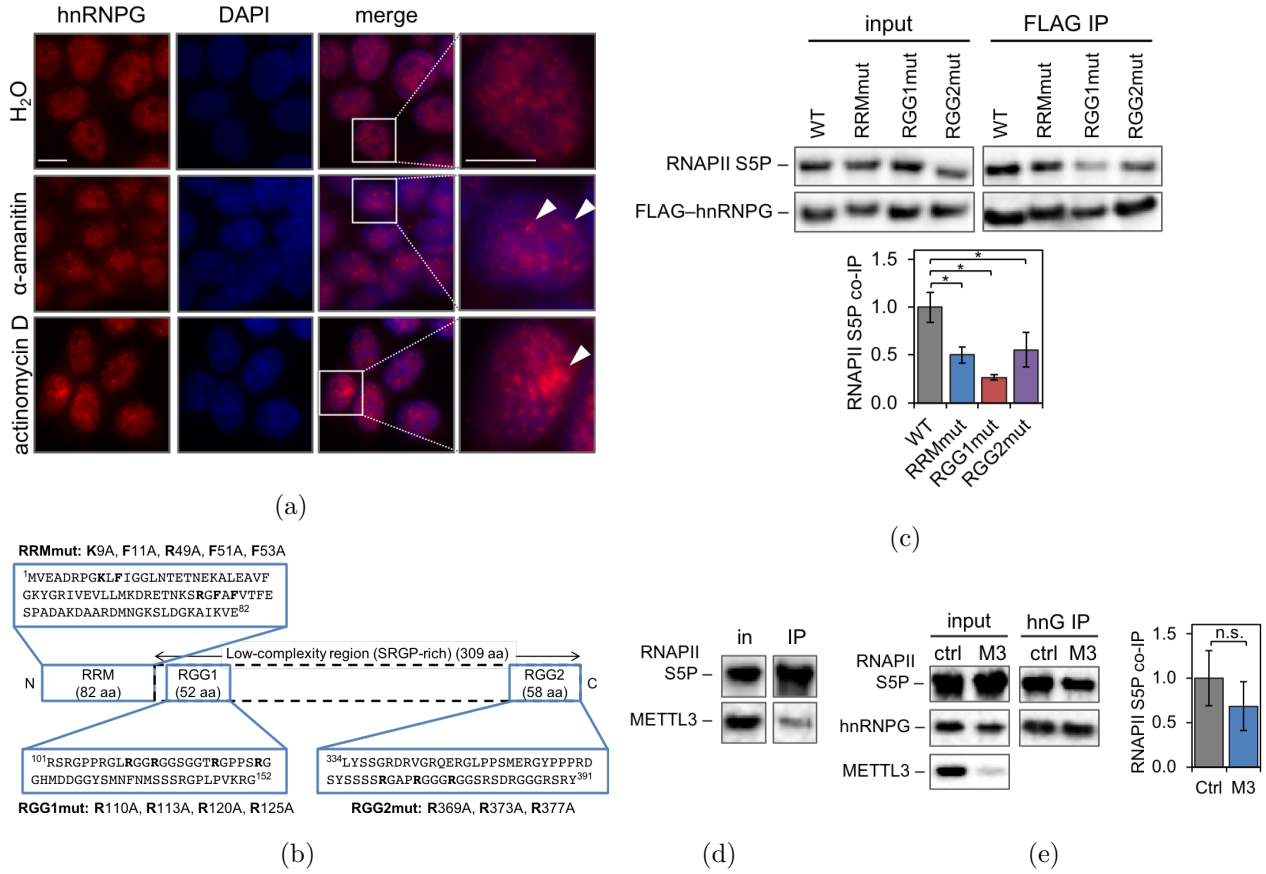


Figure 4.2: Perturbing the interaction of hnRNPG with RNA polymerase II in cells. (a) Immunofluorescence staining of hnRNPG in HEK293T cells treated with +2% v/v H_2O for 9 hours, 20 μg/mL α-amanitin for 9 hours, or 5 μg/mL actinomycin D for 2 hours. Arrowheads: dense clusters of hnRNPG in nuclei of cells treated with α-amanitin or actinomycin D. Scale bar: 10 μm. (b) Diagram showing the RRM, RGG1, RGG2, and low-complexity regions of full-length hnRNPG, as well as the mutations introduced in the RRM, RGG1, and RGG2 regions to generate RRMmut, RGG1mut, and RGG2mut. (c) Western blot showing co-IP of RNAPII S5P with FLAG–hnRNPG in chromatin extracts of HEK293T cells transfected hnRNPG siRNA and pCMV3-Flag-RBMX plasmid (WT, RRMmut, RGG1mut, or RGG2mut). input, chromatin extract. (d) Western blot showing co-IP of METTL3 with RNAPII S5P in the chromatin extract of HEK293T cells. in: input (chromatin extract); IP: RNAPII S5P IP. (e) Western blot showing co-IP of RNAPII S5P with hnRNPG in chromatin extracts of HEK293T cells transfected with control siRNA (ctrl) or METTL3 siRNA (M3). input, chromatin extract; hnG IP, hnRNPG IP. Bar graphs: RNAPII co-IP was measured relative to RNAPII co-IP for control cells (controls: c, WT; e, control siRNA). Error bars, $\pm$ 1 standard deviation; n = 3–4 biological replicates; n.s., not significant; * $p < 0.05$ by two-tailed t-test.

4.2.2 The RGG regions of hnRNPG mediate a direct interaction with the phosphorylated CTD of RNAPII

The CTD of RNAPII is the docking site for many RNA processing factors and has been shown to interact with the low-complexity regions of several RNA-binding proteins [10]. We

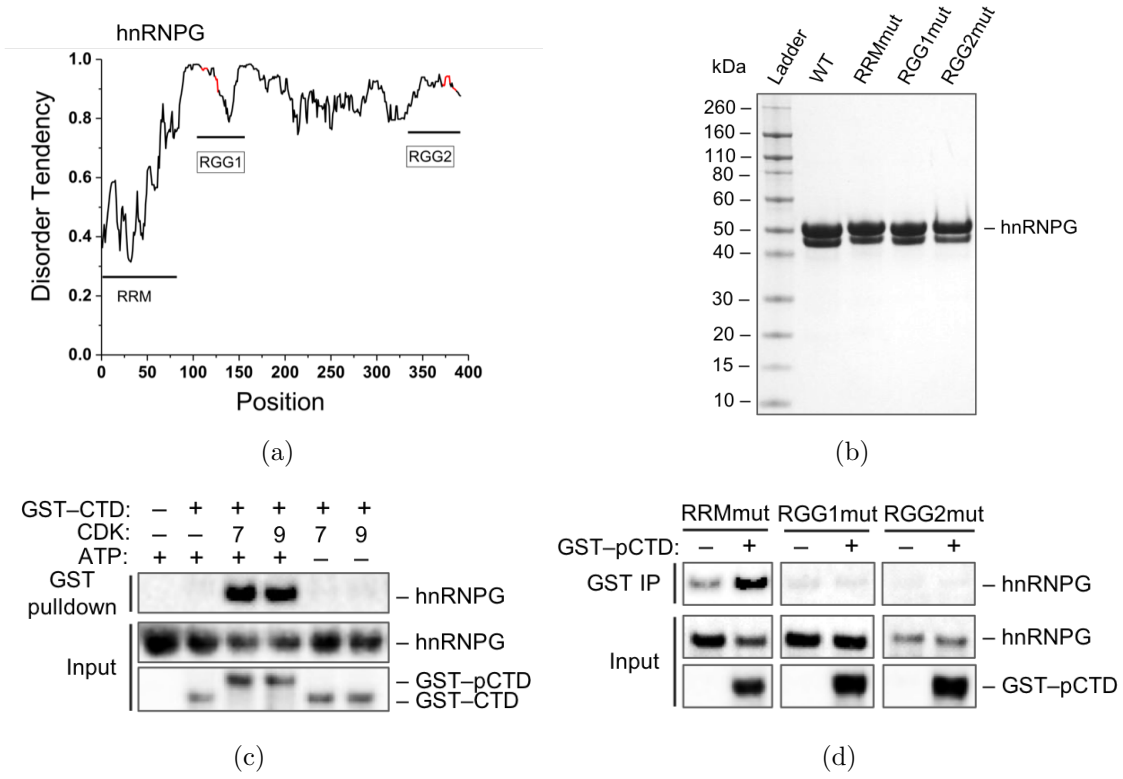


Figure 4.3: Direct interaction of hnRNPG with the RNAPII CTD *in vitro*. (a) Prediction of intrinsic disorder tendency based on the primary sequence of full-length human hnRNPG [262, 263]. The RRM, RGG1, and RGG2 regions are identified by horizontal bars. Red color of the disorder tendency curve identifies the locations of Arg-Gly-Gly (RGG) sequences. (b) Denaturing protein gel showing the full-length WT, RRMmut, RGG1mut, and RGG2mut forms of the hnRNPG protein purified from insect cells. The protein runs as two bands due to variable *N*-linked glycosylation in insect cells. Ladder: Novex Sharp Pre-stained Protein Standard (LC5800, Thermo Fisher). (c) Western blot showing pull-down of hnRNPG with the GST-tagged RNAPII CTD (GST–CTD) with or without pre-treatment of the GST–CTD with CDK7 or CDK9, with or without addition of ATP. GST–pCTD: phosphorylated GST–CTD. (d) Western blot showing co-immunoprecipitation of recombinant hnRNPG mutants RRMmut, RGG1mut, and RGG2mut with GST–CTD pre-phosphorylated with CDK7 (GST–pCTD) in vitro.

examined whether hnRNPG, which contains an extensive disordered low-complexity region (Figure 4.2(b) and 4.3(a)), can interact directly with the CTD. We used the GST-tagged CTD of RNAPII (GST–CTD) to pull down recombinant full-length hnRNPG protein purified from insect cells (Figure 4.3(b)). Phosphorylated GST–CTD was prepared by treatment with either cyclin-dependent kinase 7 (CDK7) or CDK9 prior to the pull-down. Only the phosphorylated forms of the GST–CTD were able to pull down hnRNPG (Figure 4.3(c)). Addition of kinase alone without adenosine triphosphate (ATP) did not result in hnRNPG pull-down (Figure 4.3(c)), indicating that hnRNPG interacted directly with the phosphorylated GST–CTD, rather than indirectly through the CDK7 and CDK9 proteins. To further investigate how hnRNPG interacts with the CTD, we purified the full-length hnRNPG mutants RRMmut, RGG1mut, and RGG2mut (Figure 4.3(b)). Mutations in the RRM region did not affect the interaction of hnRNPG with GST–CTD, whereas mutations in either the RGG1 or RGG2 region abolished this interaction (Figure 4.3(d)). Thus, the RGG regions are required for the direct interaction of hnRNPG with the phosphorylated CTD of RNAPII in vitro.

4.2.3 hnRNPG can interact with both RNA and the RNAPII CTD simultaneously

Next, we investigated the effect of RNA on hnRNPG binding to the phosphorylated CTD of RNAPII. Previous work has shown that hnRNPG preferentially binds to the m^6A-containing form of a 34-nucleotide hairpin derived from the long noncoding RNA metastasis-associated lung adenocarcinoma transcript 1 (*MALAT1*) [3]. We found that pre-binding hnRNPG to the m^6A-containing (methylated) or nonmethylated form of this hairpin did not influence the pull-down of hnRNPG by GST–CTD (Figure 4.4(a)). To examine whether hnRNPG binding to the CTD impacted RNA binding, we used surface plasmon resonance (SPR) to measure hnRNPG binding to immobilized *MALAT1* hairpin RNA, in the presence of increasing amounts of a peptide consisting of two heptapeptide repeats with phosphorylated

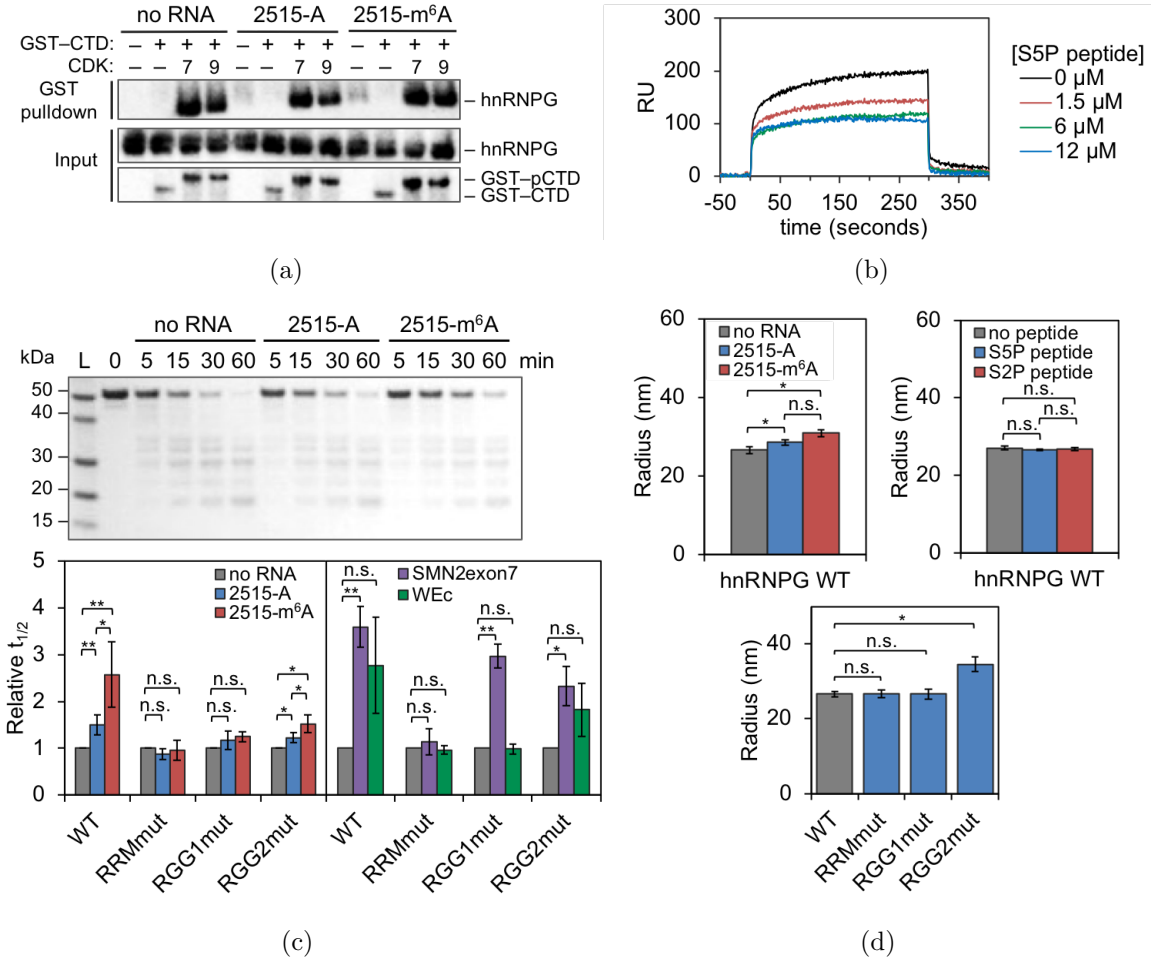


Figure 4.4: Interactions of hnRNPG with the RNAPII CTD and RNA, and hnRNPG assembly, *in vitro*. (a) Western blot showing pull-down of hnRNPG ± 2 515-A or 2 515-m^{6}A RNA with the GST–CTD ± phosphorylation by CDK7 or CDK9. GST–pCTD: phosphorylated GST–CTD. (b) Surface plasmon resonance showing binding of 3 μM hnRNPG to immobilized 2 515-A-Biotin RNA in the presence of 0–12 μM S5P peptide. RU = response units. (c) Denaturing protein gel showing time points from the limited proteolysis of wild-type hnRNPG ± 2 515-A or 2 515-m^{6}A RNA by proteinase K at 4 °C. Quantification of half-life ($t_{1/2}$) of wild-type and mutant hnRNPG proteolysis by proteinase K, with or without pre-binding of hnRNPG to equimolar amounts of 2 515-A, 2 515-m^{6}A, SMN2exon7, or WEc RNA. Half-lives are normalized to $t_{1/2}$ of wild-type or mutant hnRNPG without addition of RNA. Error bars: ± 1 standard deviation; n = 3–5 replicates; n.s., not significant by two-tailed t-test; * $p < 0.05$, ** $p < 0.01$ by two-tailed t-test. (d) Radius of wild-type and mutant hnRNPG assemblies, with or without pre-binding of hnRNPG to 2 515-A or 2 515-m^{6}A RNA, or to S5P or S2P peptide, measured by dynamic light scattering at 4 °C. Error bars: ± 1 standard deviation; n = 3–4 replicates; n.s., not significant by two-tailed t-test; * $p < 0.05$ by two-tailed t-test.

Ser5 from the RNAPII CTD (S5P peptide). The hnRNPG protein alone bound to the *MALAT1* hairpin, and increasing amounts of the S5P peptide slightly decreased but did not fully eliminate binding, even at four-fold molar concentrations relative to hnRNPG protein (Figure 4.4(b)). These results demonstrate that the interactions of hnRNPG with RNA and with the RNAPII CTD do not significantly interfere with one another.

We further investigated the roles of the different regions of hnRNPG in RNA binding using a limited proteolysis assay. hnRNPG with or without equimolar amounts of RNA was treated with proteinase K, and the degradation of full-length hnRNPG was tracked over time (Figure 4.4(c)). The *MALAT1* hairpin protected full-length hnRNPG from proteolysis, with the m^6A-containing *MALAT1* hairpin having a stronger protective effect than the non-methylated hairpin. However, the protective effect of the *MALAT1* hairpin was markedly blunted for the hnRNPG mutants RGG1mut and RGG2mut, and was abolished for RRM-mut. Mutations in the RRM region also abolished the protective effect of two other RNAs previously shown to be bound by hnRNPG: SMN2exon7 RNA and WEc RNA. Mutations in the RGG1 and RGG2 regions only slightly decreased the protective effect of SMN2exon7 RNA, which is predicted to be unstructured and is bound by the RRM region of hnRNPG [261]. In contrast, RGG region mutations markedly blunted the protective effect of WEc RNA, which is a structured RNA bound by the RGG2 region [249].

Since low-complexity regions can promote the self-assembly of proteins into higher-order structures [183, 219, 224], we also tested whether hnRNPG can self-assemble *in vitro*. In fact, purified hnRNPG formed large assemblies with a radius of ∼27 nm, corresponding to clusters of ∼180 hnRNPG proteins, as measured by dynamic light scattering (Figure 4.4(d)). The addition of nonmethylated or methylated *MALAT1* RNA, or of S5P or S2P peptides from the CTD, had minimal effects on the size of hnRNPG assemblies (Figure 4.4(d)). Mutations in the RRM or RGG1 regions did not affect the size of hnRNPG assemblies, although mutations in the RGG2 region slightly increased their size (Figure 4.4(d)). The ability of hnRNPG to self-assemble suggests that a single large hnRNPG assembly could provide a multivalent

binding surface for interactions with both nascent RNA and the phosphorylated CTD of transcribing RNAPII.

4.2.4 The RRM, RGG1, and RGG2 regions function in the regulation of gene expression by hnRNPG

To examine the effect of the RRM, RGG1, and RGG2 regions on the cellular functions of hnRNPG, we knocked down endogenous hnRNPG in HEK293T cells and overexpressed either negative control vector coding only for a FLAG tag (NCV), wild-type FLAG–hnRNPG (WT), or mutant FLAG–hnRNPG with mutations in either the RRM, RGG1, or RGG2 re-

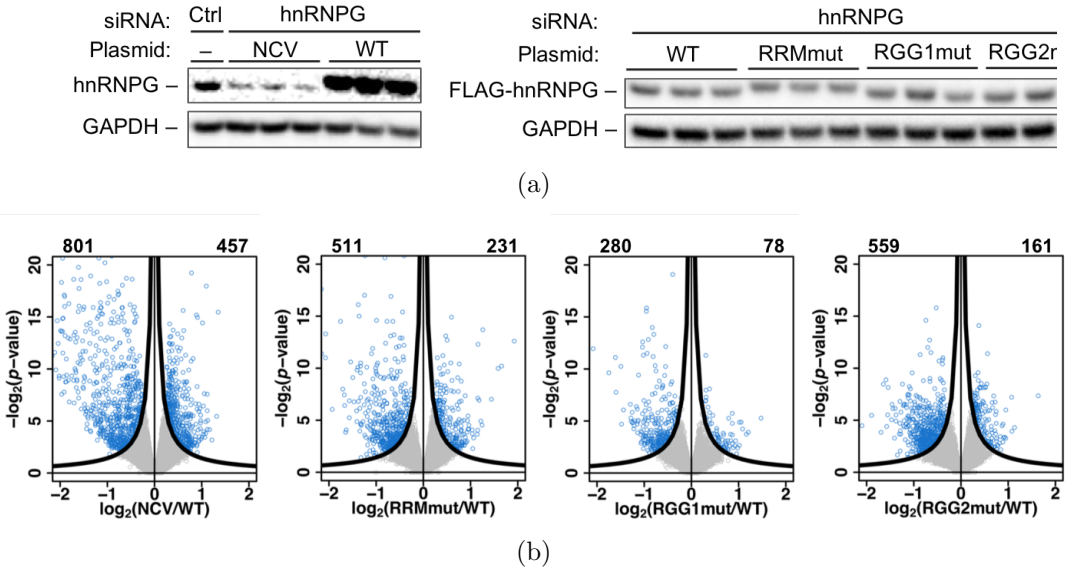
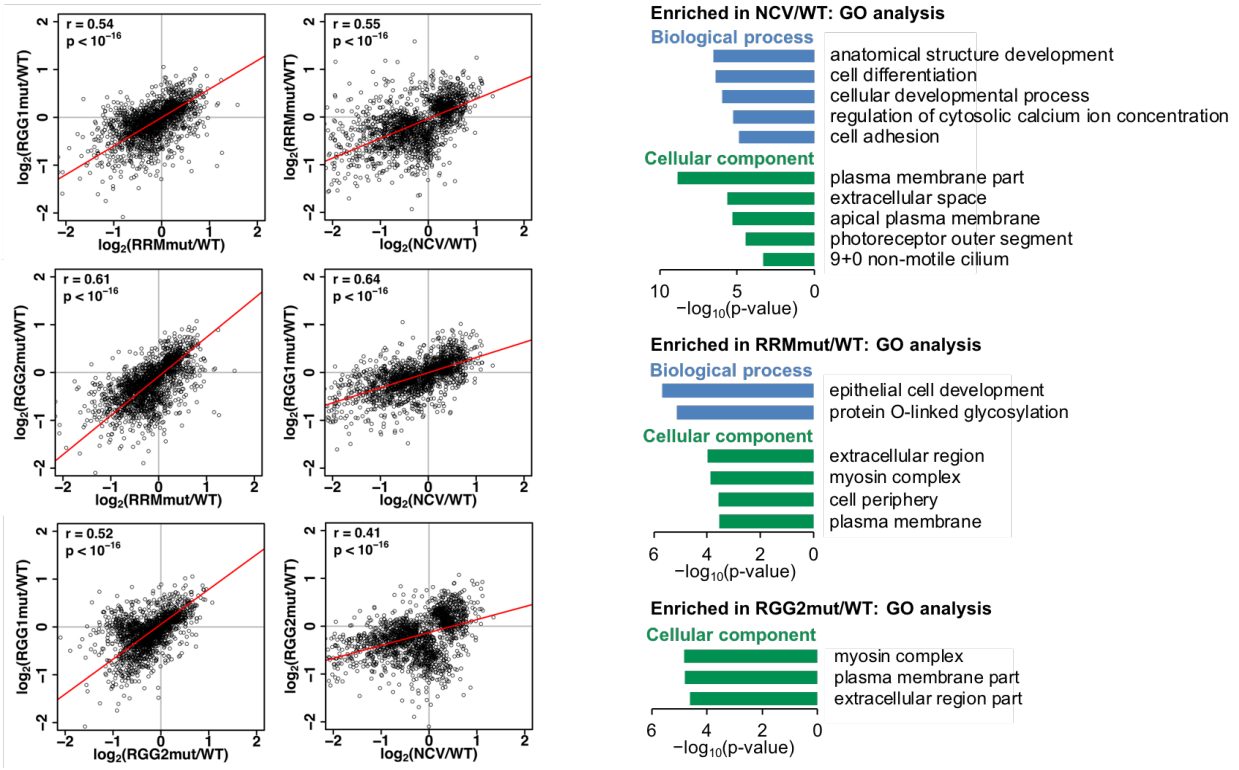


Figure 4.5: The RRM, RGG1, and RGG2 regions in the regulation of gene expression by hnRNPG. (a) Western blot showing knockdown of endogenous hnRNPG and transfection with pCMV3-Flag negative control vector (NCV) or pCMV3-Flag-RBMX encoding wild-type (WT) or mutant (RRMmut, RGG1mut, or RGG2mut) hnRNPG in whole cell lysates from HEK293T cells. Ctrl: control siRNA; GAPDH: loading control. (b) Volcano plots showing $log_2(fold\ change)$ and $-log_2(p\text{-}value)$ for changes in gene expression in NCV, RRMmut, RGG1mut, and RGG2mut relative to WT. Genes satisfying the π *value* threshold ($\mid \pi\ value \mid = \mid log_2(fold\ change) \cdot log_{10}(p\text{-}value) \mid \geq 0.4292$) were considered as differentially expressed. The number of down- and up-regulated genes are listed at the top left and top right sides, respectively, of each plot. Black curves, π *value* threshold; blue points, differentially expressed exons; gray points, non-differentially expressed exons.

gion (RRMmut, RGG1mut, or RGG2mut) (Figure 4.2(b) and 4.5(a)). We conducted mRNA sequencing (mRNA-seq) to detect changes in gene expression relative to cells expressing wild-type FLAG–hnRNPG (WT). Relative to cells expressing wild-type FLAG–hnRNPG, 1 258 genes were differentially expressed in cells expressing NCV, and 300–800 genes were differentially expressed for each of the hnRNPG mutants (Figure 4.5(b)). Moreover, changes in gene expression were correlated among the three mutants, and between each mutant and NCV (Figure 4.6(a)). This result indicates that the RRM, RGG1, and RGG2 regions all contribute to the regulation of gene expression, and that the mutations we introduced in



(a) (b)

Figure 4.6: Correlations and enrichment analysis for the regulation of gene expression by hnRNPG. (a) Correlated changes in gene expression, quantified as $log_2(fold\ change)$ relative to WT, in mRNA sequencing data for NCV, RRMmut, RGG1mut, and RGG2mut. Each point is a differentially expressed gene. r, Pearson correlation coefficient; p, p-value using Fisher transformation; red line, model II major axis linear regression. (b) Gene ontology (GO) analysis showing the $-log_{10}(p\text{-}value)$ for biological processes (blue) and cellular components (green) enriched among genes that were differentially expressed relative to WT.

these regions lead to partial loss-of-function of hnRNPG. Gene ontology analysis revealed that differentially expressed genes were enriched for functions in developmental processes (Figure 4.6(b)), consistent with the known functions of hnRNPG in neural and muscle development [247, 264].

4.2.5 The RRM, RGG1, and RGG2 regions function in the regulation of alternative splicing by hnRNPG

Previous work showed that knockdown of hnRNPG and of the m^6A methyltransferase components METTL3 and METTL14 led to correlated changes in exon splicing [3]. Using published sequencing data [3], we identified exons that were differentially expressed upon hnRNPG, METTL3, or METTL14 knockdown (regulated exons), or that showed correlated changes in expression upon hnRNPG knockdown and either METTL3 or METTL14 knockdown (co-regulated exons) (Table 4.1 and Figure 4.7(a)). Exons that were regulated by hnRNPG or co-regulated by hnRNPG and either METTL3 or METTL14 were enriched in genes that function in metabolic processes and RNA binding (Figure 4.7(b)–4.7(c)). Previous work showed that many exons that were co-regulated by hnRNPG and m^6A methyltransferase occurred in genes containing hnRNPG-bound m^6A sites [3]. Using published photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) followed by methylated RNA immunoprecipitation (MeRIP) sequencing data [3], we examined the distribution of hnRNPG-bound m^6A sites around the $3'$ and $5'$ splice sites of exons

	(co-)down-regulated exons	(co-)down-regulated exons
hnRNPG KD	21 141	18,365
METTL3 KD	30 147	24 043
METTL14 KD	22 171	19 052
hnRNPG KD and METTL3/L14 KD	11 996	9 818

Table 4.1: Differentially expressed exons upon hnRNPG, METTL3, or METTL14 KD

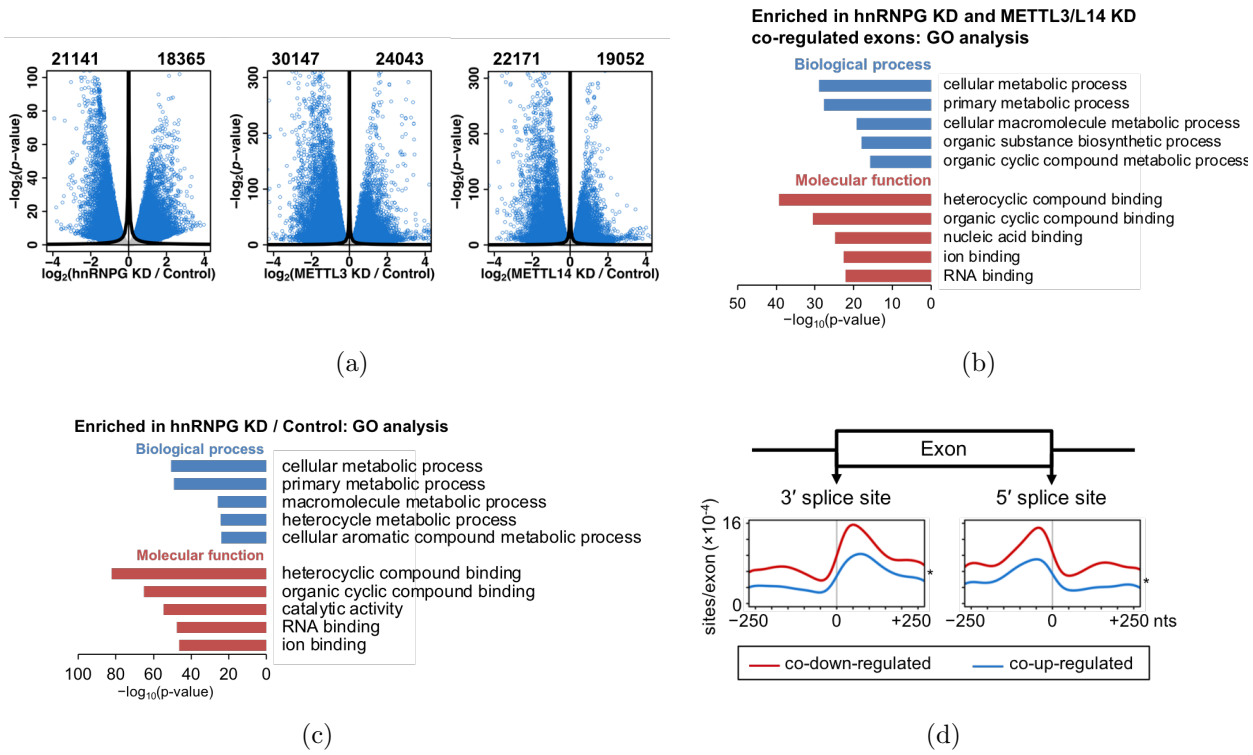


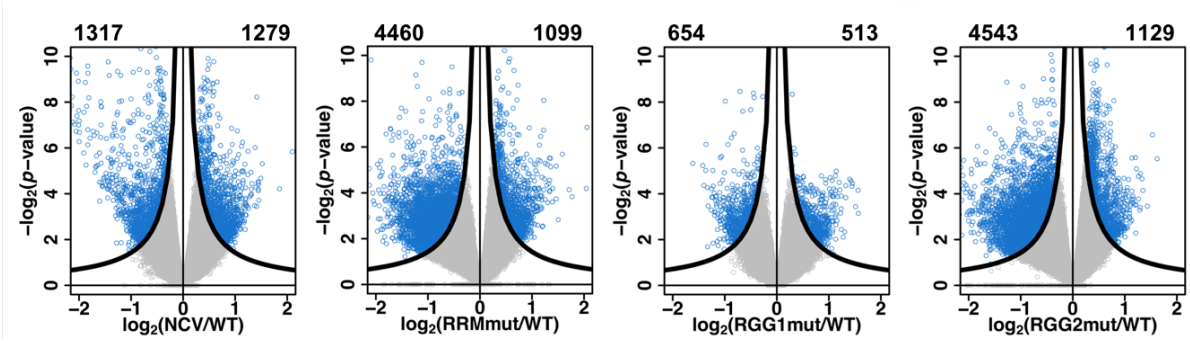
Figure 4.7: Co-regulation of exons upon hnRNPG and m^{6}A methyltransferase knockdown. (a) Volcano plots showing $log_2(fold\ change)$ and $-log_2(p\text{-}value)$ for changes in exon abundance in hnRNPG knockdown (KD), METTL3 KD, and METTL14 KD relative to their corresponding Control knockdowns (Control). Exons satisfying the π *value* threshold ($\mid \pi\ value \mid = \mid log_2(fold\ change) \cdot log_{10}(p\text{-}value) \mid \geq 0.4292$) were considered as differentially expressed exons. The number of down- and up-regulated differentially expressed exons are listed at the top left and top right sides, respectively, of each plot. Black curves, π *value* threshold; blue points, differentially expressed exons; gray points, non-differentially expressed exons. Gene ontology (GO) analysis showing the $-log_{10}(p\text{-}value)$ for biological processes (blue) and molecular functions (red) enriched among genes containing differentially expressed exons that were co-regulated by hnRNPG KD and either METTL3 KD or METTL14 KD (b), and among genes containing exons that were differentially expressed upon hnRNPG KD relative to Control KD (c). (d) Number of hnRNPG-bound m^{6}A sites per regulated exon at each site in the -250 to $+250$ nucleotide region around the $3'$ and $5'$ splice sites of exons co-down-regulated (red) or co-up-regulated (blue) upon hnRNPG KD and either METTL3 KD or METTL14 KD. * $p < 10^{-16}$ based on paired t-test between curves for down-regulated versus up-regulated exons.

	co-down-regulated exons	co-up-regulated exons
total number of exons	11 996	9 818
exons with hnRNPG-bound m^6A site at $3'$ ss $\pm$ 300 nt	4 328 (36.1%)	2 193 (22.3%)
exons with hnRNPG-bound m^6A site at $5'$ ss $\pm$ 300 nt	4 369 (36.4%)	2 062 (21.0%)

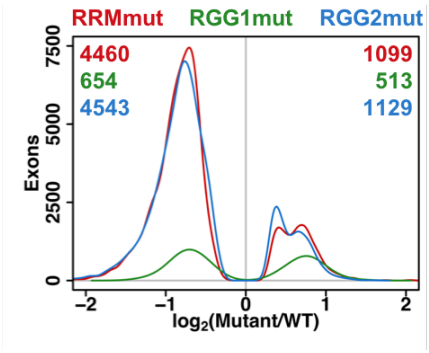
Table 4.2: hnRNPG-bound m^6A sites near splice sites of exons co-regulated by hnRNPG KD and METTL3/L14 KD

co-down- or co-up-regulated by hnRNPG and METTL3 or METTL14 (Figure 4.7(d)). We found that hnRNPG-bound m^6A sites were enriched in exonic regions near the $3'$ and $5'$ splice sites of both co-down- and co-up-regulated exons. However, co-down-regulated exons were more likely to have hnRNPG-bound m^6A sites near their splice sites, compared to co-up-regulated exons (Table 4.2).

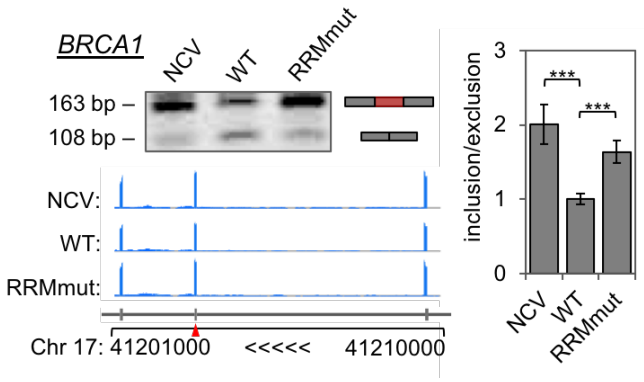
We also analyzed our mRNA-seq data for changes in exon splicing in cells expressing NCV, RRMmut, RGG1mut, or RGG2mut relative to cells expressing wild-type FLAG–hnRNPG (WT). We found thousands of differentially expressed exons (regulated exons) for each mutant and for NCV (Figure 4.8(a)–4.8(b)), and changes in exon splicing were validated by RT–PCR (Figure 4.8(c)). Notably, mutations in the RGG1 region affected the expression of fewer exons, compared to mutations in the RRM or RGG2 region. Moreover, mutations in the RRM and RGG2 regions led to many more down-regulated than up-regulated exons, suggesting that these regions predominantly function to promote exon inclusion (Figure 4.8(a)–4.8(b)). Changes in exon expression upon hnRNPG knockdown (NCV) correlated with the changes seen in cells expressing mutant FLAG–hnRNPG, and differential exon expression was also correlated among cells expressing RRMmut, RGG1mut, and RGG2mut (Figure 4.9(a)). The correlation between RRMmut and RGG2mut was stronger than the correlation of either mutant with RGG1mut, possibly due to overlapping functions of the RRM and RGG2 regions in nascent RNA binding.



(a)



(b)

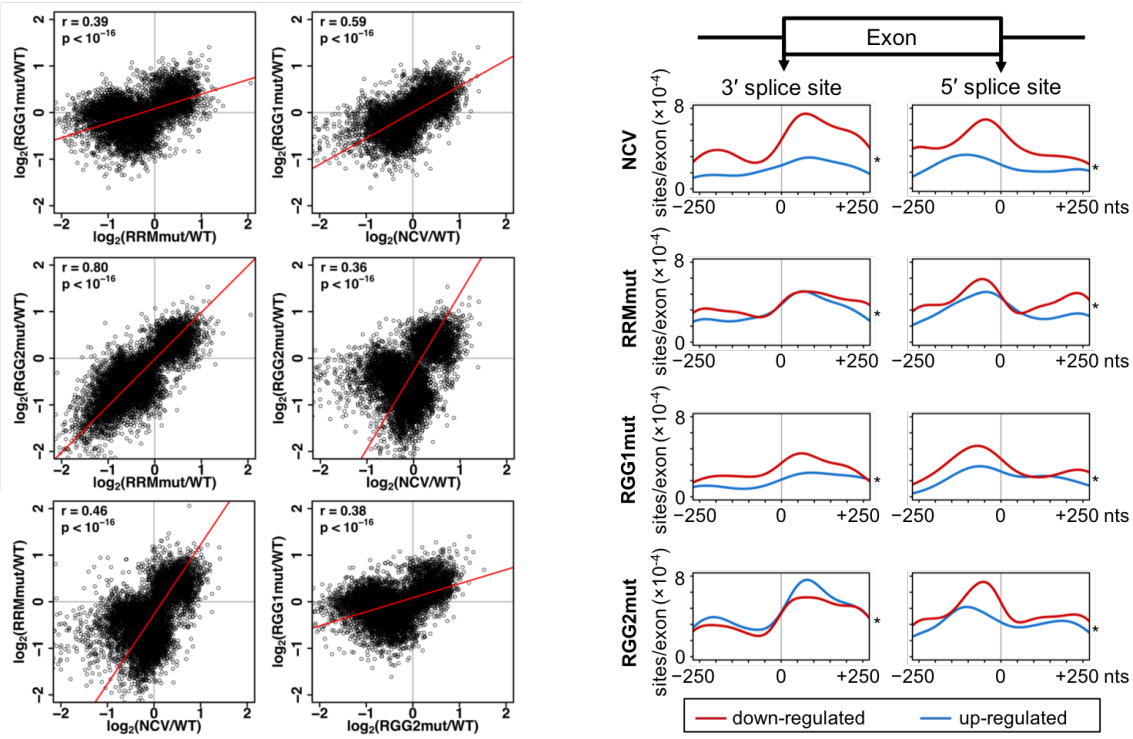


(c)

Figure 4.8: The RRM, RGG1, and RGG2 regions in the regulation of alternative splicing by hnRNPG. (a) Volcano plots showing $log_2(fold\ change)$ and $-log_2(p\text{-}value)$ for changes in exon abundance in RRMmut, RGG1mut, and RGG2mut relative to WT. Exons satisfying the π *value* threshold ($\mid \pi\ value \mid = \mid log_2(fold\ change) \cdot log_{10}(p\text{-}value) \mid \geq 0.4292$) were considered as differentially expressed exons. The number of down- and up-regulated differentially expressed exons are listed at the top left and top right sides, respectively, of each plot. Black curves, π *value* threshold; blue points, differentially expressed exons; gray points, non-differentially expressed exons. (B) Distribution of differentially expressed exons showing the number of exons at each $log_2(fold\ change)$ for RRMmut (red), RGG1mut (green), and RGG2mut (blue) relative to WT. The number of down- and up-regulated differentially expressed exons for each mutant are listed at the top left and top right corners, respectively, of the plot, in the color of the corresponding mutant. (c) Native polyacrylamide gel and quantification of RT–PCR validation, as well as mRNA-seq reads showing differential exon usage in BRCA1 RNA for NCV, WT, and RRMmut. Red arrowhead: alternatively spliced exon. Error bars: $\pm$ 1 standard deviation; $n = 3$ biological replicates; *** $p < 0.001$ by two-tailed t-test.

4.2.6 A role for m^{6}A site position in the regulation of alternative splicing by hnRNPG

To further investigate the functions of the RRM, RGG1, and RGG2 regions in the m^{6}A-dependent regulation of alternative splicing by hnRNPG, we examined the pattern of hnRNPG-bound m^{6}A sites around the splice sites of alternatively spliced exons. Consistent with the pattern around exons co-regulated by hnRNPG and m^{6}A methyltransferase (Figure 4.7(d)), hnRNPG-bound m^{6}A sites were enriched in exonic regions near the splice sites of exons



(a) (b)

Figure 4.9: Correlations and m^{6}A site positions for exons regulated by hnRNPG. (a) Correlated changes in exon expression, quantified as $log_2(fold\ change)$ relative to WT, in mRNA sequencing data for RRMmut, RGG1mut, and RGG2mut. Each point is a differentially expressed exon. r, Pearson correlation coefficient; p, p-value using Fisher transformation; red line, model II major axis linear regression. (b) Number of hnRNPG-bound m^{6}A sites per regulated exon at each site in the −250 to +250 nucleotide region around the 3′ and 5′ splice sites of exons down-regulated (red) or up-regulated (blue) in RRMmut, RGG1mut, and RGG2mut relative to WT. * $p < 10^{-16}$ based on paired t-test between curves for down-regulated versus up-regulated exons.

down-regulated upon hnRNPG knockdown (NCV), while this pattern was much weaker in up-regulated exons (Figure 4.9(b)). An enrichment of hnRNPG-bound m^6A sites in exonic regions near splice sites was also observed among exons that were differentially expressed in cells with mutations in the RRM, RGG1, or RGG2 region of hnRNPG (Figure 4.9(b)). Notably, exons regulated by RGG1mut had fewer hnRNPG-bound m^6A sites near splice sites than exons regulated by RRMmut or RGG2mut. Exons regulated by RGG2mut tended to have more hnRNPG-bound m^6A sites near splice sites than exons regulated by RRMmut, particularly downstream of the $3'$ splice site of up-regulated exons. In addition, exons down- and up-regulated by RGG2mut exhibited different patterns in their hnRNPG-bound m^6A sites: exons down-regulated by RGG2mut were enriched in hnRNPG-bound m^6A sites in exonic regions near the $5'$ splice site, whereas exons up-regulated by RGG2mut were enriched in hnRNPG-bound m^6A sites in exonic regions near the $3'$ splice site. Thus, exons whose regulation depended on the RGG2 region were the most likely to have splice-site-proximal hnRNPG-bound m^6A sites, and the position of these sites differed for down- versus up-regulated exons.

The enrichment of hnRNPG-bound m^6A sites around the splice sites of exons regulated by hnRNPG in an RGG2-dependent manner is consistent with the function of the RGG2 region in the preferential binding of m^6A-modified RNAs [3]. In addition, genes containing RGG2mut-regulated exons were enriched in metabolic functions (Figure 4.10(a)), as was also observed for exons co-regulated by hnRNPG and m^6A methyltransferase (Figure 4.7(b)). Genes containing RGG2mut-regulated exons also tended to function in nervous system development. This result is consistent with the previously reported role of hnRNPG in neural development [247, 264]. Moreover, a frameshift mutation that leads to the truncation of most of the RGG2 region of hnRNPG has been linked to an intellectual disability syndrome in humans [248]. Genes containing exons regulated by NCV, RRMmut, and RGG1mut were also enriched for functions in metabolism and development (Figure 4.10(b)–4.10(d)). As was observed for RGG2mut-regulated exons, genes containing RRMmut-regulated exons were

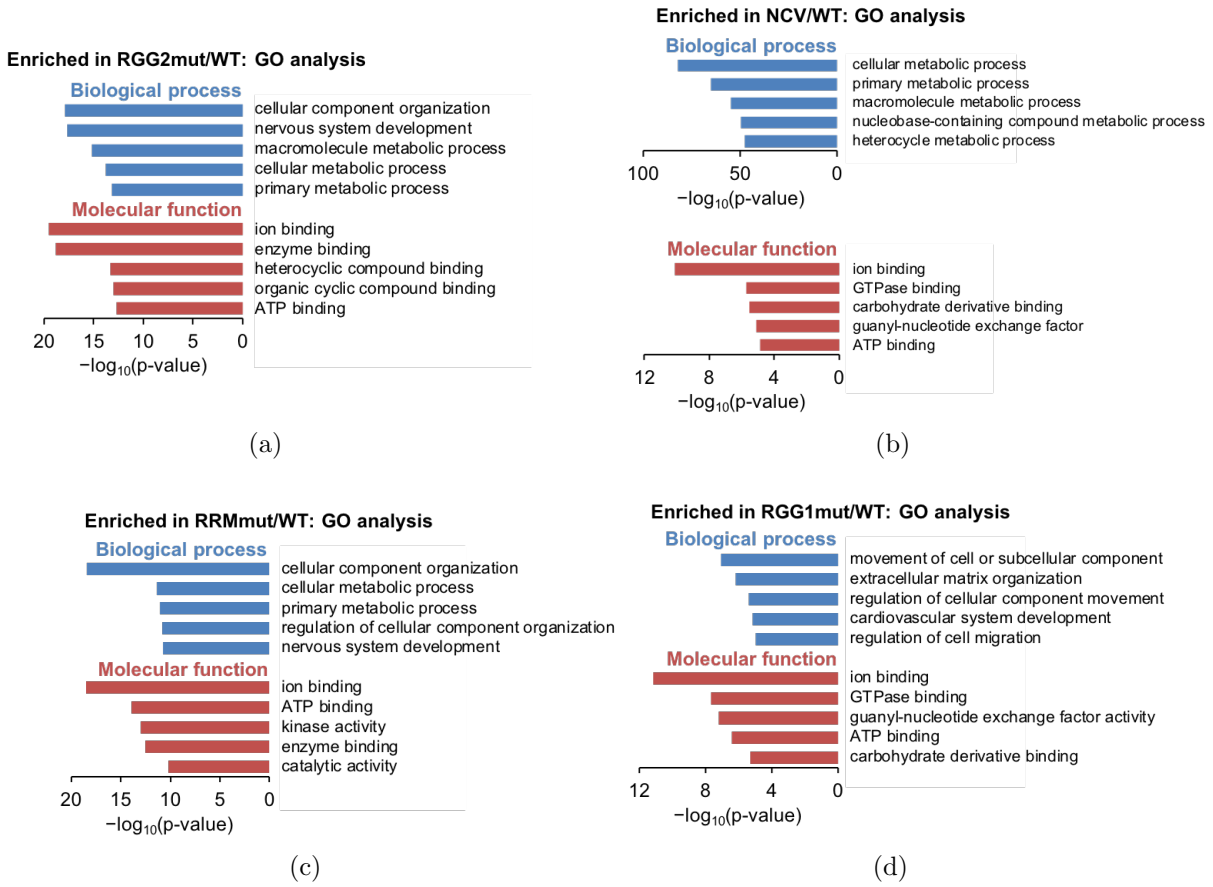
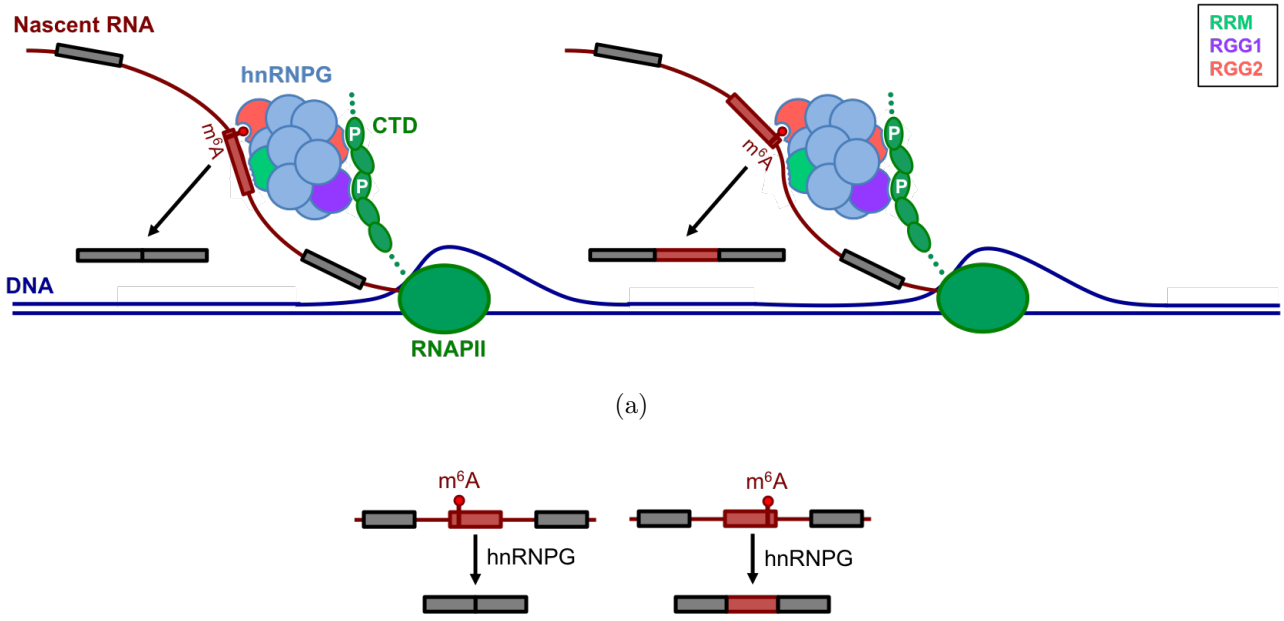


Figure 4.10: Gene ontology (GO) analysis of genes containing exons differentially expressed in RGG2mut (a), NCV (b), RRMmut (c), or RGG1mut (d) relative to WT, showing the $-log_{10}(p\text{-}value)$ for enriched biological processes (blue) and molecular functions (red).

associated with functions in nervous system development, consistent with the strong correlation between exon expression changes in RRMmut and RGG2mut (Figure 4.9(a)). On the other hand, genes containing RGG1mut-regulated exons were associated with cardiovascular system development. Interestingly, hnRNPG was previously reported to regulate a splicing event associated with dilated cardiomyopathy [265].

Based on these results, we propose a model for the co-transcriptional m^6A-dependent regulation of alternative splicing by hnRNPG (Figure 4.11). The RGG1 and RGG2 regions interact with phosphorylated repeats in the CTD of transcribing RNAPII, while the RRM and RGG2 regions interact with nascent mRNA, which contains m^6A modifications

that enhance RGG2 region binding. The RRM, RGG1, and RGG2 regions of hnRNPG all contribute to alternative splicing regulation through their interactions with RNA and with the CTD. However, exons regulated in an RGG2-dependent manner are most likely to have hnRNPG-bound m^{6}A sites near their splice sites. Moreover, for the RGG2-dependent regulation of exons by hnRNPG, the positions of hnRNPG-bound m^{6}A sites determine whether hnRNPG favors exon inclusion or exclusion.



(a)

(b)

Figure 4.11: Model for the co-transcriptional m^{6}A-dependent regulation of alternative splicing by hnRNPG. (a) hnRNPG assemblies bind both to phosphorylated (P) heptapeptide repeats of RNAPII using the RGG1 (purple) and RGG2 (salmon) regions, and to nascent RNA using the RRM (green) and RGG2 regions. The RGG2 regions preferentially bind to m^{6}A sites in the nascent RNA, which target hnRNPG to co-transcriptionally regulate the splicing of exons in an m^{6}A site position-dependent manner. Red disc: m^{6}A site; dark red box: regulated alternative exon; gray box: constitutive exon. (b) The position of m^{6}A sites in exonic regions near the $3'$ or $5'$ splice site of the regulated exon determines whether hnRNPG favors exon exclusion or inclusion. Red disc: m^{6}A site; dark red box: regulated alternative exon; gray box: constitutive exon.

4.3 Discussion

In this study, we showed that the RNA-binding protein hnRNPG uses RGG regions to directly interact with the phosphorylated CTD of transcribing RNAPII. hnRNPG binding to RNA and to the RNAPII CTD could occur simultaneously, and hnRNPG assembled into large complexes *in vitro*. In cells, the RRM, RGG1, and RGG2 regions of hnRNPG functioned in the regulation of gene expression and alternative splicing. By examining the position of hnRNPG-bound m^6A sites around the splice sites of regulated exons, we found that m^6A site position relative to exon splice sites was associated with the direction of RGG2-dependent exon regulation. Together, these results support a model in which hnRNPG assemblies interact co-transcriptionally with nascent RNA and with the RNAPII CTD for gene regulation, while m^6A sites in nascent RNA promote hnRNPG binding to modulate alternative splicing.

The CTD of RNAPII functions to coordinate co-transcriptional RNA processing with the different stages of transcription by recruiting constitutive RNA processing factors through direct interactions. However, the role of the CTD in recruiting alternative splicing factors, particularly through direct interactions, is less clear [31]. This study demonstrates that an alternative splicing factor, hnRNPG, directly interacts with the RNAPII CTD. The selective interaction of hnRNPG with the phosphorylated form of the CTD, as well as the effect of different transcription inhibitors on the interaction of hnRNPG with RNAPII, support our conclusion that hnRNPG interacts with the RNAPII CTD co-transcriptionally. We also found that the direct interaction of hnRNPG with the phosphorylated CTD depended on RGG repeats in the low-complexity region of hnRNPG. The CTD has previously been shown to interact with the low-complexity regions of other RNA-binding proteins, and these interactions were proposed to function in the regulation of transcription [10, 231]. The binding of low-complexity regions was inhibited by CTD phosphorylation [230] and promoted by RNA [183, 225]. In contrast, phosphorylation of the CTD was required for its interaction with hnRNPG, and RNA did not influence hnRNPG binding to the CTD. Thus, hnRNPG forms

a distinct type of interaction with the RNAPII CTD for the co-transcriptional regulation of gene expression and alternative splicing.

The abundant mRNA modification m^6A has been implicated in the regulation of alternative splicing, yet the mechanisms by which m^6A influences splicing are not well understood. m^6A modifications are deposited co-transcriptionally and are enriched in exonic regions near splice sites [85–88], but it is unknown whether m^6A can regulate co-transcriptional splicing, or how the positions of m^6A sites relative to splice sites can modulate alternative splicing. Alternative splicing regulation by RNA-binding proteins is highly dependent on the positions of protein binding sites [43]. Thus, the positions of m^6A sites, which determine the positions of m^6A reader protein binding sites, likely also have strong effects on alternative splicing regulation. Indeed, we found that hnRNPG-bound m^6A sites occurred in different positions relative to the splice sites of exons, depending on whether the exons were up-regulated or down-regulated by hnRNPG in an RGG2-dependent manner. Therefore, we propose that the positions of m^6A sites relative to splice sites determine the direction of exon regulation by hnRNPG. In this way, hnRNPG regulates alternative splicing co-transcriptionally and in an m^6A position-dependent manner.

The regulation of alternative splicing involves a complex 'splicing code' of *cis*- and *trans*-acting factors [43]. In this study, we showed that a *trans*-acting RNA binding protein, hnRNPG, interacts with the transcription machinery and co-transcriptionally regulates alternative splicing. In addition, a *cis*-acting RNA modification, m^6A, modulates the regulation of alternative splicing by hnRNPG by recruiting hnRNPG to specific positions relative to the regulated exon's splice sites. The mechanisms by which hnRNPG regulates alternative splicing once recruited to its site of action remain unclear. Since hnRNPG interacts directly with transcribing RNAPII, hnRNPG might modulate alternative splicing by regulating transcription rate or pausing, as well as CTD phosphorylation or assembly. Alternatively, hnRNPG might modulate splicing factor recruitment, for instance by competing for binding sites on nascent RNA or on the CTD. The m^6A-dependent regulation of alternative splicing by hn-

RNPG demonstrates how the CTD of RNAPII and m^6A modifications in nascent mRNA can act together to modulate the co-transcriptional regulation of alternative splicing by a low-complexity m^6A reader protein.

4.4 Materials and Methods

4.4.1 Cloning and purification of hnRNPG

The sequence encoding full-length human hnRNPG protein was amplified from human HeLa cDNA libraries (637203, Clontech) and subcloned into the pGEX-6p-1 vector using BamHI and XhoI restriction sites. Plasmids encoding the hnRNPG mutants RRMmut, RGG1mut, and RGG2mut were prepared by QuikChange mutagenesis (200524, Agilent) and Gibson assembly (E2611L, New England BioLabs), with the following mutations in the encoded proteins: K9A, F11A, R49A, F51A, and F53A in RRMmut; R110A, R113A, R120A, and R125A in RGG1mut; R369A, R373A, and R377A in RGG2mut. The mutant hnRNPG sequences were cloned into pCMV3-Flag-RBMX (HG16560-NF, Sino Biological) by Gibson assembly for expression in human cells. The wild-type and mutant hnRNPG sequences were cloned into the pAcGP67a vector using BamHI and NotI restriction sites for expression in insect cells. A His_8 tag was added to the N-terminus for affinity purification. A fast-folding variant of protein G, NuG2b (DTYKLVIVLNGTTFTYTTEAVDAATAEKVFKQYANDAGVDGEWTYDAATKTFTVTE [266, 267]), was added to the N-terminus to increase protein stability.

A baculovirus expression system was used for expression of proteins in High Five insect cells as previously described [268]. The secreted proteins were purified using nickel nitrilotriacetic agarose (Ni-NTA) resin (30230, Qiagen). The resin was washed with 10 mM Tris-Cl (pH 7.4), 1 M NaCl, 2.5 mM $MgCl_2$, 10% v/v glycerol buffer. The hnRNPG protein was released from the resin by cleavage C-terminal to the His_8–NuG2b tag with His-tagged PreScission Protease (Z03092, GenScript) in storage buffer (10 mM Tris-Cl (pH 7.4), 500 mM NaCl, 2.5 mM $MgCl_2$, 10% v/v glycerol) at 4 °C overnight, and the protein was concen-

trated with a 30 kDa centrifugal filter (UFC803024) and either stored at 4 °C or flash-frozen in liquid nitrogen and stored at −80 °C. Immediately before use, hnRNPG protein stocks were spun at 21 K · g at 4 °C for 10 minutes, and the supernatant was used as the new hnRNPG stock. The concentrations of hnRNPG stocks were measured by Bradford assay (23236, Thermo).

4.4.2 Cell culture and transfection

Human embryonic kidney (HEK) 293T/17 cells (CRL11268) were obtained from the American Type Culture Collection (ATCC) and cultured under standard conditions. For Figure 4.1(d), 20 nM control siRNA (1027281, Qiagen) or hnRNPG siRNA (G1: SI00700077, Qiagen; G2: SI00700084, Qiagen) was transfected into HEK293T cells using Lipofectamine RNAiMAX, and the cells were collected for chromatin extraction 62 hours after transfection. For Figure 4.1(e), HEK293T cells were treated with 0.5% v/v dimethyl sulfoxide (DMSO), 5 μg/mL actinomycin D (A9415, Sigma), or 100 μM 5,6-dichloro-1-β-D-ribofuranosylbenzimidazole (DRB) (D1916, Sigma) for 2 hours before collection of cells for chromatin extraction. For Figure 4.1(f), HEK293T cells were treated with 0.12% v/v DMSO or 6 μM camptothecin (C9911, Sigma) for 5 hours before collection of cells for chromatin extraction. For Figure 4.2(a), HEK293T cells were treated with +2% v/v H_2O or 20 μg/mL α-amanitin (A2263, Sigma) for 9 hours, or with 5 μg/mL actinomycin D (A9415, Sigma) for 2 hours, before fixation for immunofluorescence. For Figures 4.2(c) and 4.5(a), 24 ng/mL of either pCMV3-Flag negative control vector (CV016, Sino Biological) or pCMV3-Flag-RBMX plasmid (WT, RRMmut, RGG1mut, or RGG2mut), as well as 10 nM hnRNPG siRNA (SI00700077, Qiagen), were transfected into HEK293T cells using Lipofectamine 2000 (11668019, Thermo), and the cells were collected for RNA extraction 68 hours after transfection. For Figure 4.2(e), 20 nM control siRNA (1027281, Qiagen) or METTL3 siRNA (SI04317096, Qiagen) was transfected into HEK293T cells using Lipofectamine RNAiMAX (13778150, Thermo), and the cells were collected for chromatin extraction 60 hours after

transfection.

4.4.3 Immunofluorescence

HEK293T cells were grown in tissue-culture-treated 8-well slides (80826, ibidi), fixed with 4% (w/v) formaldehyde, permeabilized with phosphate-buffered saline (PBS) (10× solution: SH30258.01, HyClone) containing 0.2% v/v Triton X-100, and blocked overnight at 4 °C in blocking buffer (PBS with 2% w/v bovine serum albumin (A7030, Sigma)). After blocking, the cells were incubated with 365 ng/mL rabbit anti-hnRNPG antibody (ab190352, Abcam) in blocking buffer for 1 hour at room temperature, incubated with 1 μg/mL goat anti-rabbit IgG Alexa Fluor 647 antibody (A-31573, Thermo) in blocking buffer for 1 hour, stained with PBS containing 0.1 μg/mL 4′,6-diamidino-2-phenylindole (DAPI) (D1306, Thermo) for 2 minutes, and covered with ibidi mounting medium (50001, ibidi). Between steps, the cells were washed 2–3 times with PBS for 5 minutes per wash. The slides were imaged on an Olympus DSU spinning disk confocal microscope at the University of Chicago Integrated Light Microscopy Core Facility.

4.4.4 Preparation of cell extracts

HEK293T cells were washed and detached from the cell culture plate with PBS, and then pelleted at 500 · g for 3 minutes. For extraction of whole cell lysate, the cells were resuspended in whole cell lysis buffer: 300 mM NaCl, 100 mM Tris-Cl (pH 8), 0.2 mM ethylenediaminetetraacetic acid (EDTA), 0.1% v/v Triton X-100, and 10% v/v glycerol supplemented with freshly added 1% v/v protease inhibitor (25955-11, Nacalai USA) and 1% v/v phosphatase inhibitor (07575-51, Nacalai USA). After rotating at 4 °C for 30 minutes, cell debris was pelleted at 16 K · g at 4 °C for 5 minutes, and the supernatant whole cell lysate was collected and stored at −20 °C.

Cell fractionation and chromatin extraction were performed based on published protocols [269–271]. HEK293T cells were resuspended in cytoplasmic extraction buffer: 10 mM Tris-

Cl (pH 7.4), 10 mM KCl, and 0.1% v/v Triton X-100 supplemented with freshly added 1% v/v protease and phosphatase inhibitors. After incubating on ice for 20 minutes, nuclei were pelleted at 12 K $\cdot$ g at 4 °C for 10 minutes. The pellet was washed once with cytoplasmic extraction buffer. The nuclei were resuspended with nuclear extraction buffer: 10 mM Tris-Cl (pH 7.4), 0.2 mM $MgCl_2$, and 1% v/v Triton X-100 supplemented with freshly added 1% v/v protease and phosphatase inhibitors. After incubating on ice for 15 minutes, chromatin was pelleted at 12 K $\cdot$ g at 4 °C for 10 minutes. The chromatin pellets were resuspended in 5 U/mL micrococcal nuclease (N3755, Sigma), 20 mM Tris-Cl pH 7.4, 100 mM KCl, 2 mM $MgCl_2$, 1–3 mM $CaCl_2$, 0.3 M sucrose, and 0.1% v/v Triton X-100 supplemented with freshly added 1% v/v protease and phosphatase inhibitors. After rotating at 4 °C for 1 hour, the digestion reaction was stopped by adding 5 mM EDTA. After centrifuging at 2 K $\cdot$ g at 4 °C for 5 minutes, the chromatin extract was collected as the supernatant.

4.4.5 Immunoprecipitation

For immunoprecipitation, HEK293T whole cell lysate or chromatin extract was combined with 28 μg/mL rabbit anti-DDDDK antibody (ab1162, Abcam), 20 μg/mL rabbit anti-hnRNPG antibody (Ab #1 or not specified: ab190352, Abcam), 1/20 v/v rabbit anti-hnRNPG antibody (Ab #2: 14794, Cell Signaling Technology; Ab #3: NBP2-34152, Novus Biologicals), 20 μg/mL rabbit isotype control antibody (ab199376, Abcam) in a total volume of 90–100 μL per 100-mm plate of cells. After rotating at 4 °C overnight, 1.5 mg of protein A or protein G Dynabeads (10002D or 10004D, Thermo) were added. After rotating at 4 °C for 2 hours, the tubes were placed on a magnetic separation rack, and the supernatant unbound fraction was collected and stored at −20 °C. The beads were washed four times with 200 μL of wash buffer (300 mM NaCl, 100 mM Tris-Cl (pH 8), 0.2 mM EDTA, and 0.1% v/v Triton X-100). Finally, 30 μL of 4× LDS sample buffer (NP0008, Thermo) were added, and the tubes were incubated at 95 °C for 5 minutes and then placed on a magnetic rack. The supernatant immunoprecipitate (IP) was collected and stored at −20 °C. The

concentrations of the input and unbound fractions were measured by Bradford assay. For Western blotting, 5 μg of input (whole cell lysate or chromatin extract) or unbound fraction was combined with 1× LDS and 250 mM dithiothreitol (DTT), and 5 μL (one sixth) of the IP fraction was combined with 250 mM DTT.

4.4.6 GST–CTD pull-down

To phosphorylate the GST–CTD, 500 ng of recombinant GST-tagged human RNAPII CTD (SRP2120, Sigma) were combined with 240 ng of CDK7–Cyclin H–MNAT1 (PV3868, Thermo) or CDK9–Cyclin T1 (14-685, Sigma) in 20 μL of 8 mM 3-(N-morpholino)propanesulfonic acid (MOPS) (pH 7), 0.2 mM EDTA, and 1 mM $MgCl_2$ supplemented with freshly added 0.1 mM ATP and 0.25 mM DTT, and then incubated at 30 °C for 1 hour. Control reactions without GST–CTD, without kinase, or without ATP were incubated under the same conditions. For the GST–CTD pull-down and immunoprecipitation assays, this 20-μL reaction was combined with 60 μL of wash buffer and 100 pmol of hnRNPG in 20 μL of storage buffer (10 mM Tris-Cl (pH 7.4), 500 mM NaCl, 2.5 mM $MgCl_2$, 10% v/v glycerol), and then rotated at 4 °C overnight in a total volume of 100 μL. One tenth (10 μL) of the binding mixture was taken as the input. A 20-μL volume of glutathione magnetic agarose beads (78601, Thermo) for the pull-down (Figures 4.3(c) and 4.4(a)), or GST mouse monoclonal antibody magnetic bead conjugate (11847, Cell Signaling Technology) for the immunoprecipitation (Figure 4.3(d)), was blocked in 100 μL of 1 mg/mL BSA in wash buffer at 4 °C for 1 hour, resuspended in 20 μL of wash buffer, and added to the remaining 90 μL of the binding mixture. After rotating at 4 °C for 2 hours, the tubes were placed on a magnetic rack, and the supernatant unbound fraction was collected. The beads were washed four times with 200 μL of wash buffer. For elution, the beads were resuspended in 30 μL 50 mM reduced glutathione in wash buffer and incubated at 22 °C, 300 rpm for 20 minutes for the pull-down, or resuspended in 30 μL of 4× LDS and boiled at 95 °C for 5 minutes for the immunoprecipitation. The tubes were placed on a magnetic rack, and the supernatant eluate was collected and stored at −20 °C.

For blotting, 10 μL of the input fraction (one tenth of total input) or 10 μL of the eluate (one third) were prepared in solutions with final concentrations of 1–2× LDS and 250 mM DTT. The pull-down samples were visualized by Western blotting for hnRNPG (below), while the immunoprecipitation samples were visualized by Sypro Ruby protein blot staining (S-11791, Thermo) according to the manufacturer's instructions. Both the pull-down and immunoprecipitation samples were also visualized by Western blotting for GST (below).

4.4.7 Western blotting

All samples were incubated at 95 °C for 10 minutes, separated on a 4–12% polyacrylamide Bis-Tris protein gel (NP0336BOX, Thermo), and transferred to polyvinylidene fluoride membranes (IPVH00010, Millipore). The membranes were blocked in 20 mM Tris-Cl (pH 7.6), 150 mM NaCl, and 0.1% v/v Triton X-100, with either 5% w/v bovine serum albumin for RNAPII NTD blots, or milk (170-6404, Bio-Rad) for all other blots. The blots were probed with 0.7 μg/mL anti-RNAPII S5P antibody (ab5131, Abcam), 1 μg/mL anti-RNAPII S2P H5 antibody (920204, BioLegend), 0.365 μg/mL anti-hnRNPG antibody (ab190352, Abcam), 1 μg/mL anti-DDDDK antibody (ab1162, Abcam), 0.23 μg/mL anti-METTL3 antibody (15073-1-AP, Proteintech), 1/1000 v/v anti-GST antibody (2624, Cell Signaling Technology), or 0.05 μg/mL anti-GAPDH antibody (A00192-40, Genscript), followed by 0.5 μg/mL mouse anti-rabbit light chain (ab99697, Abcam), 0.03 μg/mL goat anti-rabbit IgG (ab97051, Abcam), 1/2000 v/v rat anti-mouse light chain (ab99632, Abcam), or 0.05 μg/mL goat anti-mouse IgG antibody conjugated to horseradish peroxidase (ab97023, Abcam). For quantification of co-immunoprecipitation, Western blot bands were quantified using ImageLab software, and RNAPII bands were normalized to hnRNPG bands and to input as follows: $(RNAPII_{IP}/hnRNPG_{IP})/(RNAPII_{input}/hnRNPG_{input})$, where X_Y is the intensity of band X in lane Y.

4.4.8 Oligonucleotides

The following RNA oligonucleotides were synthesized and purified by high-performance liquid chromatography and/or denaturing polyacrylamide gel electrophoresis, as previously described [239].

2 515-A: 5′- AAUGUGAAGGACUUUCGUAACGGAAGUAAUUCAA

2 515-m^{6}A: 5′- AAUGUGAAGGm^{6}ACUUUCGUAACGGAAGUAAUUCAA

2 515-A-Biotin: 5′- AAUGUGAAGGACUUUCGUAACGGAAGUAAUUCAA-Biotin

2 515-m^{6}A-Biotin: 5′- AAUGUGAAGGm^{6}ACUUUCGUAACGGAAGUAAUUCAA-Biotin

The following DNA oligonucleotides were ordered from Integrated DNA Technologies, purified by denaturing polyacrylamide gel electrophoresis, annealed in 50 mM Tris-Cl (pH 7.5), 500 mM KCl at 94 °C for 1 minute, and incubated at room temperature for 3–5 minutes. The resulting DNA templates were used to *in vitro* transcribe the SMNexon7 and WEc RNAs with T7 RNA polymerase (M0251, New England BioLabs), and the RNA transcripts were precipitated and purified by denaturing gel electrophoresis.

SMNexon7-Fw: 5′-TAATACGACTCACTATAGGTTTTAGACAAAATCAAAAAGAAGG
AAGGTGCTCACATTCCTTAAATTAAGGA

SMNexon7-Rv: 5′-TCCTTAATTTAAGGAATGTGAGCACCTTCCTTCTTTTTGATTTT
GTCTAAAACCTATAGTGAGTCGTATTA

WEc-Fw: 5′-TAATACGACTCACTATAGGGTACGACGGATATCGTGGGGGGGGA
AATTGCTTTCGGTTCCGACTCTG

WEc-Rv: 5′-CAGAGTCGGAACCGAAAGCAATTTCCCCCCCCCACGATATCCGTC
GTACCCTATAGTGAGTCGTATTA

The following DNA oligonucleotides were ordered from Integrated DNA Technologies, purified by denaturing polyacrylamide gel electrophoresis, and used for RT–PCR validation of splicing changes detected by mRNA sequencing:

BRCA1_Fw: 5′-GTGGTCAATGGAAGAAACCACC

BRCA1_Rv: 5′-CCTTCACCACAGAAGCACCAC

4.4.9 Surface plasmon resonance

Surface plasmon resonance data were collected on a BioRad ProteOn XPR36 instrument at the University of Chicago BioPhysics Core Facility. The RNA oligonucleotides 2 515-A-Biotin and 2 515-m^{6}A-Biotin were immobilized on separate lanes of an NLC sensor chip (1765021 Bio-Rad). Recombinant hnRNPG protein was flowed over the chip at a concentration of 3 μM with 0–12 μM S5P peptide (ab18488, Abcam) in filtered and degassed 10 mM 4-(2-hydroxyethyl)piperazine-1-ethanesulfonic acid (HEPES) (pH 8.6), 500 mM NaCl, 3 mM EDTA, and 0.05% v/v Polysorbate 20 buffer.

4.4.10 Limited proteolysis

The RNA was diluted to a concentration of 25 μM in 30 mM Tris-Cl (pH 7.4) and incubated at 90 °C for 1 minute, and then at room temperature for 3–5 minutes. For limited proteolysis, 200 pmol of wild-type or mutant hnRNPG were combined 200 pmol of RNA in 14 mM Tris-Cl (pH 7.4), 400 mM NaCl, 2 mM $MgCl_2$, and 8% v/v glycerol in a total volume of 40 μL, and the mixture was incubated for 10 minutes at room temperature, and then placed on ice. Proteinase K (EO0491, Thermo) was added for a ratio of 1:700 w/w proteinase:hnRNPG, and the reaction was incubated on ice. Time points were taken at 5, 15, 30, and 60 minutes by combining 9.21 μL of the reaction with 1× stop solution (1× LDS, 1% v/v protease inhibitor) and incubating at 95 °C for 5 minutes before placing on dry ice. The 0-minute time point was prepared by combining 1.7 μg of hnRNPG with 1× stop solution. Immediately after the 60-minute time point was collected, all of the samples were separated on a 4–12% polyacrylamide Bis-Tris gel and stained with PageBlue Protein Staining Solution (24620, Thermo) according to the manufacturer's instructions. Bands were quantified using ImageJ, and the fraction of full-length hnRNPG in each lane was calculated and fit to a single exponential to derive the half-life ($t_{1/2}$).

4.4.11 Dynamic light scattering

For Figure 4.4(d), 9 μM wild-type or mutant hnRNPG was prepared in 14 mM Tris-Cl (pH 7.4), 400 mM NaCl, 2 mM $MgCl_2$, and 8% v/v glycerol, with or without 7.2 μM RNA, for a total volume of 40 μL; or 9 μM wild-type hnRNPG was prepared in storage buffer (10 mM Tris-Cl (pH 7.4), 500 mM NaCl, 2.5 mM $MgCl_2$, 10% v/v glycerol) with or without 36 μM S5P or S2P peptide (ab18488 and ab12793, Abcam), for a total volume of 40 μL. The mixture was incubated for 10 minutes at room temperature, and then placed on ice. Light scattering measurements were collected at 4 °C on a Wyatt DynaPro NanoStar instrument and analyzed using the DYNAMICS software at the University of Chicago BioPhysics Core Facility. Each measurement was the average of ten 20-second acquisitions. Each acquisition was fit to an auto-correlation curve using the regularization algorithm, with upper and lower cut-offs of 107 μs and 1.5 μs. Acquisitions with an auto-correlation curve baseline outside the 0.99–1.01 range or error sum-of-squares >100 were filtered out before averaging.

4.4.12 mRNA sequencing

HEK293T cells were transfected with hnRNPG siRNA and pCMV3-Flag (NCV) or pCMV3-Flag-RBMX (WT, RRMmut, RGG1mut, or RGG2mut) as described above, with three 100-mm-plate replicates for each of the five different plasmid transfections. Total RNA was extracted from HEK293T cells by Trizol (15596026, Thermo) extraction followed by an additional chloroform extraction. Libraries were prepared using the Tru-Seq Stranded mRNA LT Sample Prep Kit (RS-122-9005DOC, Illumina). The libraries were sequenced twice, and each time the 15 libraries were sequenced in two lanes of one flow cell by Illumina HiSeq 4000 with paired-end 100-bp reads at the University of Chicago Genomics Facility. The reads from both sequencing runs were combined for the analysis.

4.4.13 RT–PCR

Total RNA, collected for mRNA sequencing as described above, was reverse transcribed using the SuperScript III First-Strand Synthesis System (18080-051, Thermo), and then amplified using Taq DNA Polymerase (EP0401, Thermo) under the following conditions: 95 °C for 3 minutes, 30 cycles of [95 °C for 30 seconds, 50 °C for 30 seconds, 72 °C for 1 minute], and 72 °C for 10 minutes. The PCR products were resolved on a 10% polyacrylamide Novex TBE gel (EC62762BOX, Thermo) and stained with SYBR Gold Nucleic Acid Gel Stain (S11494, Thermo). Exon inclusion was quantified using ImageJ.

4.4.14 mRNA sequencing analysis

The mRNA sequencing data from this study (NCV, WT, RRMmut, RGG1mut, RGG2mut) and from a previous study (hnRNPG knockdown, METTL3 knockdown, METTL14 knockdown, and their corresponding Control knockdowns [3]) were aligned to the human genome version hg19 using the STAR computer program version 2.5.3a [272] with soft-clipping, yielding approximately 96 million (standard deviation: 9.4 million) mapped reads per sample from this study, and approximately 132 million (standard deviation: 38 million) mapped reads per sample from the previous study. Gene and exon boundaries from RefSeq [273] were extracted from the University of California Santa Cruz (UCSC) table browser [274] for human genome version hg19. Differential expression levels of exons and genes in RefSeq were analyzed using FeatureCount [275] followed by DEseq2 [276]. The fold change (FC) in gene and exon expression levels was calculated as $log_2(Counts_{[NCV\ or\ mut]}/Counts_{WT})$ for NCV, RRMmut, RGG1mut, and RGG2mut, or as $log_2(Counts_{KD}/Counts_{Control})$ for hnRNPG, METTL3, and METTL14 knockdowns (KD). Biological relevance (fold change) and statistical significance were considered simultaneously via the $\pi\ value = -log_2(FC) \cdot log_{10}(p\text{-}value)$ [277]. Differentially expressed genes and exons were selected using the threshold $|\pi\ value| \geq 0.4292$, which corresponds to significance level $\alpha = 0.1$ under assumption of independence between fold change and p-value (volcano plots show broad range of p-values

for a given fold change, and a broad range in fold change for a given p-value). Co-down- or co-up-regulated exons were exons that were differentially expressed, with $log_2(FC) < 0$ or $log_2(FC) > 0$, respectively, both upon hnRNPG KD and upon either METTL3 KD or METTL14 KD. Differentially spliced exons were identified as differentially expressed exons (DEE) for which $log_2(FC_{DEE}) - log_2(FC_{nonDEE}) > 0.9 \cdot log_2(FC_{DEE})$ for some non-differentially expressed exon (nonDEE) in the same gene. Based on this analysis, the majority of differentially expressed exons were also differentially spliced exons, and differentially spliced exons showed the same patterns of down- versus up-regulation as differentially expressed exons: RRMmut sequencing yielded 3 191 down-regulated and 894 up-regulated differentially spliced exons, RGG1mut yielded 503 down- and 433 up-regulated differentially spliced exons, and RGG2mut yielded 3 574 down- and 944 up-regulated differentially spliced exons relative to WT. For correlation plots comparing changes in gene or exon expression in different samples, genes or exons that were differentially expressed in either NCV, RRMmut, RGG1mut, or RGG2mut based on the $|\pi\ value| \geq 0.4292$ threshold were combined. The union of differentially expressed genes or exons was plotted to compare the $log_2(FC)$ in two different sequencing samples relative to WT. The Pearson correlation coefficient r, and the associated p-value based on the Fisher transformation, were calculated using R statistical software. Model II major axis linear regression was performed using lmodel2 in R statistical software. Gene ontology (GO) analysis was performed using the enrichment analysis tool provided by the Gene Ontology Consortium [278–280].

4.4.15 Analysis of hnRNPG-bound m^{6}A site distribution

Reads from previously published photoactivatable ribonucleoside-enhanced crosslinking and immunoprecipitation (PAR-CLIP) followed by methylated RNA immunoprecipitation (MeRIP) sequencing data [3] were aligned to the human genome version hg19 using the STAR computer program version 2.5.3a [272] with soft-clipping, yielding approximately 17.7 million reads (standard deviation: 15.2 million) for each sample. The hnRNPG-bound m^{6}A sites

were identified based on two criteria: PAR-CLIP T-to-C mutation profiles uncovered using PARalyzer version 1.1 [255], and positive enrichment of m^6A-immunoprecipitated fragment counts relative to input [60] at RRACH sites found in hg19 reference chromosomal sequences (direct search for RRACH sequence motif for sites on + strand, and direct search for DGTYY motif for sites on − strand). For the analysis of the distribution of hnRNPG-bound m^6A sites around the splice sites of differentially regulated exons, we identified all the hnRNPG-bound m^6A sites with position x relative to the splice site (defined as position 0) such that $-300 \leq x \leq +300$ nucleotides, for either the $3'$ or $5'$ splice sites of any of the differentially regulated exons in the selected category. For instance, for the red curve in Figure 4.7(d), we identified all hnRNPG-bound m^6A sites that occurred within 300 nucleotides of the $3'$ splice site of any of the exons co-down-regulated upon hnRNPG KD and either METTL3 KD or METTL14 KD. For each category of differentially expressed exons, the distribution of hnRNPG-bound m^6A sites around each splice site was converted by kernel density estimation using R statistical software. The resulting density at each position was multiplied by the constant n/m, where n is the total number of hnRNPG-bound m^6A sites within 300 nucleotides of the splice site, and m is the total number of differentially expressed exons in the selected category. The resulting value at each position corresponds to the mean number of hnRNPG-bound m^6A sites per exon at that position, among all differentially expressed exons in the selected category. These values (sites/exon) were plotted at each position x, with $-250 \leq x \leq +250$, in Figures 4.7(d) and 4.9(b). Rapid identification of overlapping genomic features was made possible by BEDTools intersect function version 2.17.0 [281].

4.4.16 Data Deposition

The sequencing data have been deposited to National Center for Biotechnology Information Gene Expression Omnibus database under accession number GSE114311.

Chapter 5

Pseudouridine Modifications Have Context-Dependent Mutation and Stop Rates in High-Throughput Sequencing

Acknowledgement: This chapter is derived from an article published in *RNA Biology* by Taylor & Francis [5]. The authors of that article were: Katherine I. Zhou, Wesley C. Clark, David W. Pan, Matthew J. Eckwahl, Qing Dai, and Tao Pan. Author contributions: Conceptualization, K.I.Z., W.C.C., and T.P.; Methodology, K.I.Z., W.C.C., and T.P.; Software, W.C.C., D.W.P., and M.J.E.; Formal Analysis, K.I.Z. and W.C.C.; Investigation, K.I.Z.; Writing – Original Draft, K.I.Z.; Writing – Review & Editing, K.I.Z. and T.P.; Supervision, T.P.

5.1 Introduction

The abundant RNA modification pseudouridine (Ψ) has been mapped transcriptome-wide by chemically modifying pseudouridines with carbodiimide and detecting the resulting reverse transcription stops in high-throughput sequencing. However, these methods have limited sensitivity and specificity. Only a small fraction of the newly discovered pseudouridine sites were reproducibly detected by more than one pseudouridine sequencing study [184]. In addition, quantitative mass spectrometry estimates the Ψ/U ratio in mammalian mRNAs at $\sim$0.2–0.4%, which would correspond to thousands of pseudouridine sites, or several times more sites than have been identified even with the pre-enrichment of pseudouridylated RNAs [179]. One element that limits the specificity and sensitivity of current methods is the use of reverse transcription stops to map pseudouridine sites: stops can also result from RNA structure or degradation, and pseudouridine sites that are close to another downstream pseudouridine or to the $3'$ end of an RNA molecule are less likely to be detected. An alternative approach would be to detect mutations introduced through misincorporation during reverse transcription, rather than detecting stops. Mutation-based approaches have

been used to map RNA modifications transcriptome-wide [62, 71, 282, 283], and the combined analysis of both stop and mutation rates can improve the detection of modification sites as compared to the analysis of either stop or mutation rate alone [188]. Since sequence context can influence the identities of misincorporated bases [188], accounting for the surrounding nucleotide sequence could further improve the detection of RNA modification sites.

In this work, we show that pseudouridine modifications in CMC-treated human ribosomal RNA have context-dependent mutation and stop rates in high-throughput sequencing libraries prepared under specific reverse transcription conditions. By testing different reverse transcriptase enzymes and divalent cations, we found conditions in which reverse transcriptase reads through CMC-modified pseudouridine in a synthetic RNA oligo. We then used these reverse transcription conditions to prepare high-throughput sequencing libraries from CMC-treated and mock-treated human rRNA. The proportions of stops and mutations observed at pseudouridine sites in CMC-treated RNA varied depending on the identity of the nucleotide $3'$ to pseudouridine. By defining different stop and mutation thresholds for different sequence contexts, we were able to improve the sensitivity and specificity of pseudouridine detection relative to context-independent thresholds. Thus, accounting for context-dependent stop and mutation rates can enhance the detection of RNA modification sites in high-throughput sequencing data.

5.2 Results

5.2.1 Reverse transcription through CMC-modified pseudouridine in an RNA oligo

We set out to find conditions in which reverse transcription can bypass CMC-modified pseudouridine (CMC-Ψ). We synthesized a 27-nucleotide RNA oligo containing a single pseudouridine, treated it with CMC to obtain an oligo containing a single CMC-Ψ, and reverse transcribed both the CMC-Ψ-containing and untreated Ψ-containing oligos under various

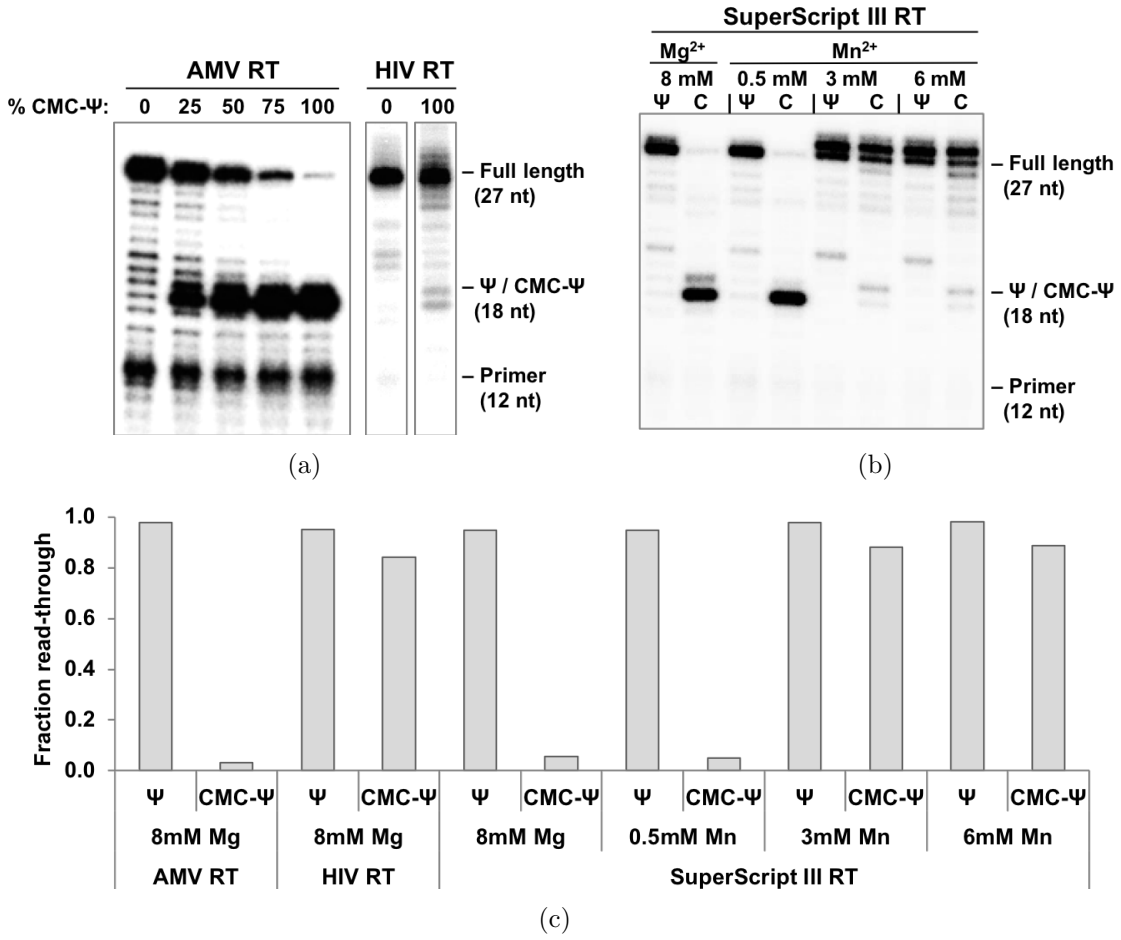


Figure 5.1: Reverse transcription through CMC-modified Ψ. Reverse transcription of a synthetic RNA oligo (Oligo Ψa) containing Ψ or CMC-Ψ (C) with AMV RT or HIV RT (a), or with SuperScript III RT under varying divalent cation conditions (b). RNA oligo sequence: 5′-UACACUCAGXUCGGACUAAAGCUGCUC (X = Ψ or CMC-Ψ). (c) Quantification of Ψ or CMC-Ψ read-through by different reverse transcriptase enzymes under varying divalent cation conditions.

conditions. Avian myeloblastosis virus (AMV) reverse transcriptase (RT) synthesized full-length complementary DNA (cDNA) from the Ψ-containing oligo template with high efficiency (Figures 5.1(a) and 5.1(c)), as expected given that pseudouridine and uridine are essentially indistinguishable by RT enzymes. When the CMC-Ψ-containing oligo was used as the template, most of the cDNAs terminated one nucleotide 3′ to the CMC-Ψ. When varying proportions of the Ψ-containing and CMC-Ψ-containing oligo templates were reverse transcribed, the quantity of truncated cDNAs scaled with the proportion of CMC-Ψ-containing

template. Since human immunodeficiency virus (HIV) RT has previously been shown to read through bulky adducts in template RNAs [284]. we tested whether HIV RT could read through CMC-modified pseudouridine. Indeed, HIV RT produced nearly equal amounts of full-length cDNA from Ψ-containing and CMC-Ψ-containing oligo templates, corresponding to ∼84% read-through of CMC-Ψ (Figures 5.1(a) and 5.1(c)). Since the manganese divalent cation (Mn^{2+}) has previously been used to enhance read-through of bulky $2'$-O-adducts by SuperScript II RT [282], we investigated the effect of different divalent cations on CMC-Ψ read-through. While SuperScript III RT behaved similarly to AMV RT under standard reaction conditions with divalent cation Mg^{2+}, replacing Mg^{2+} with at least 3 mM Mn^{2+} (with 0.5 mM of each dNTP) led to ∼88% read-through of CMC-Ψ (Figure 5.1(b)–5.1(c)). Thus, we identified two conditions that facilitate reverse transcription through CMC-Ψ: (1) reverse transcription with HIV RT and (2) reverse transcription with SuperScript III RT in the presence of 3 mM Mn^{2+}.

5.2.2 High-throughput sequencing of pseudouridine sites in human ribosomal RNA

Since the Watson–Crick face of pseudouridine is indistinguishable from that of uridine, reverse transcriptases incorporate an adenosine (A) nucleotide opposite of Ψ. However, CMC modifies the Watson–Crick face of pseudouridine, so reverse transcriptases that read through CMC-Ψ might incorporate T, C, or G rather than A opposite of CMC-Ψ. If misincorporation occurs, then CMC-modified pseudouridine sites can be identified as mutated U's in RNA sequencing data. To test this possibility, we prepared cDNA libraries of human rRNA using reverse transcription conditions that allow CMC-Ψ read-through (Figure 5.2(a)). We isolated total RNA from human embryonic kidney 293T cells (HEK293T) and size-selected for RNA molecules over 200 nucleotides long to enrich for rRNA. After fragmentation, a control RNA oligo with one partially modified site (50% Ψ and 50% U) was added. Next, the RNA was split into two parts: one part was treated with CMC followed by reversal at al-

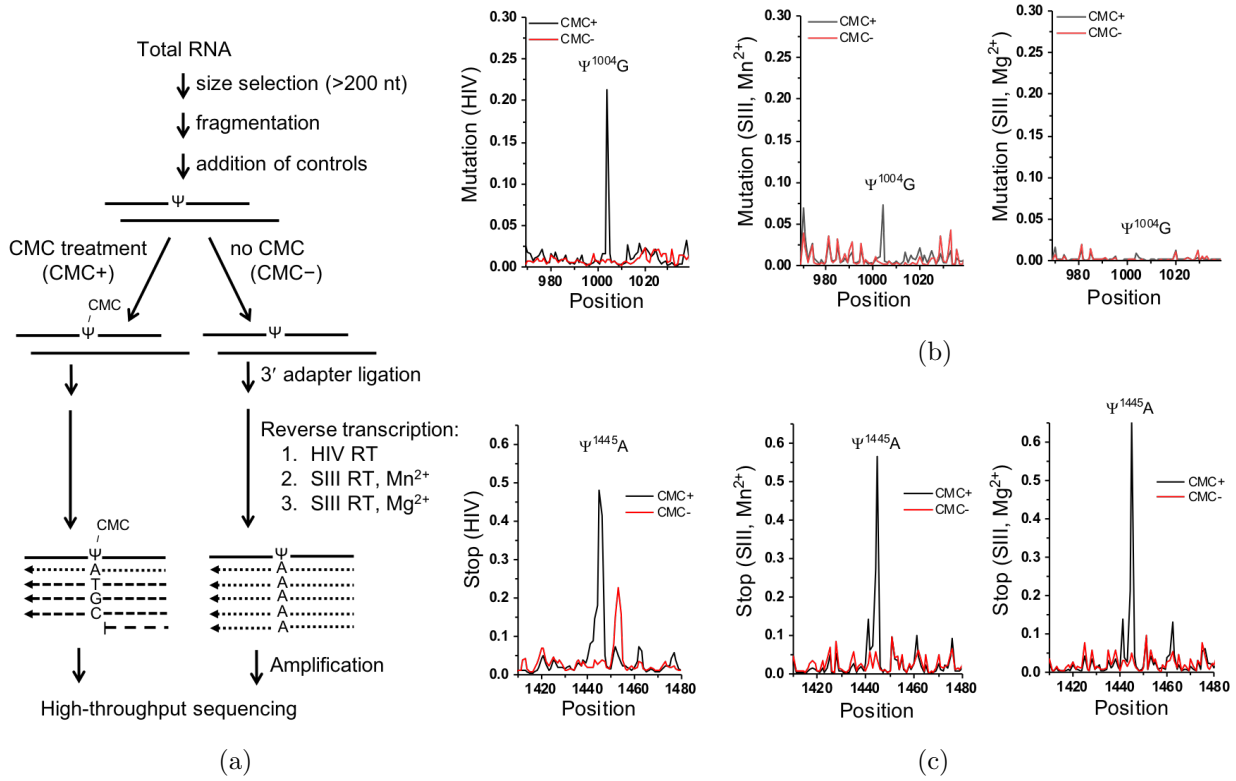


Figure 5.2: High-throughput sequencing of Ψ sites in human rRNA. (a) Sequencing workflow for detection and analysis of Ψ sites in human rRNA. Reverse transcription through CMC-modified Ψ sites with HIV RT or SuperScript III (SIII) RT can produce read-through cDNAs with or without mismatches, or truncated cDNAs. (b) Mutation rates around Ψ1004 in18S rRNA, which has a +1 G context, in libraries prepared with HIV RT, SuperScript III RT with Mn^{2+}, or SuperScript III RT with Mg^{2+}. (c) Stop rates around Ψ1445 in 18S rRNA, which has a +1 A context, in libraries prepared with HIV RT, SuperScript III RT with Mn^{2+}, or SuperScript III RT with Mg^{2+}.

kaline pH (CMC+), while the other part was mock-treated in buffers lacking CMC (CMC−). Following end repair and 3′ adapter ligation, the CMC+ and CMC− samples were each split into three equal parts that were reverse transcribed with (1) HIV RT, (2) SuperScript III RT with Mn^{2+} (SIII RT, Mn^{2+}) or (3) SuperScript III RT with Mg^{2+} (SIII RT, Mg^{2+}). For the CMC-treated sample (CMC+), reverse transcription conditions (1) and (2) were expected to generate cDNAs with misincorporations at CMC-Ψ sites, while condition (3) was expected to generate cDNAs terminating one nucleotide 3′ to CMC-Ψ. The cDNAs generated from all six conditions (CMC+/−, with 3 reverse transcription conditions each) were gel-purified,

amplified, and sequenced.

We evaluated the mutation and stop rates at known pseudouridine sites in CMC-treated and mock-treated human 18S and 28S rRNAs [285]. To minimize the background mutation rate, one nucleotide was trimmed from either end of each read. Stop count assignments were also shifted over by one nucleotide position to account for this end-trimming step. When we took the difference (Δ) between mutation or stop rates in the CMC+ and CMC− samples, we found that known pseudouridine sites in rRNA tended to have higher Δmutation rates and lower Δstop rates in the sequencing libraries prepared using HIV RT or SIII RT, Mn^{2+} than in libraries prepared using SIII RT, Mg^{2+} (Figure 5.3). Although these general trends were consistent with our expectations, the Δmutation rates at known pseudouridine sites were unexpectedly low, with a median Δmutation rate < 0.04 even in the libraries prepared under reverse transcription conditions that favored CMC-Ψ read-through. Moreover, the Δstop rates in libraries prepared using HIV RT or SIII RT, Mn^{2+} were only slightly lower than the Δstop rate in libraries prepared using SIII RT, Mg^{2+}, suggesting that CMC-Ψ read-through during library preparation was less efficient than the CMC-Ψ read-through we observed during reverse transcription of our model oligo (Figure 5.1(a)–5.1(b)). To test whether

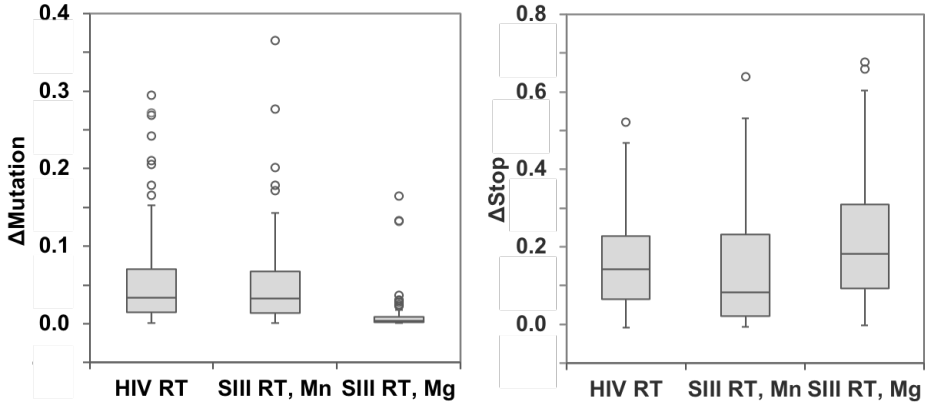
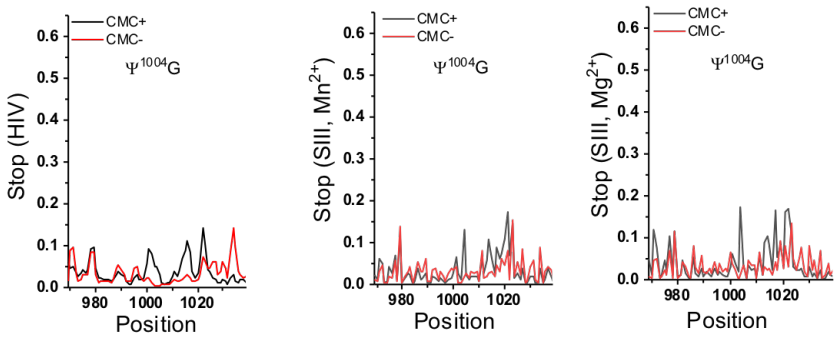
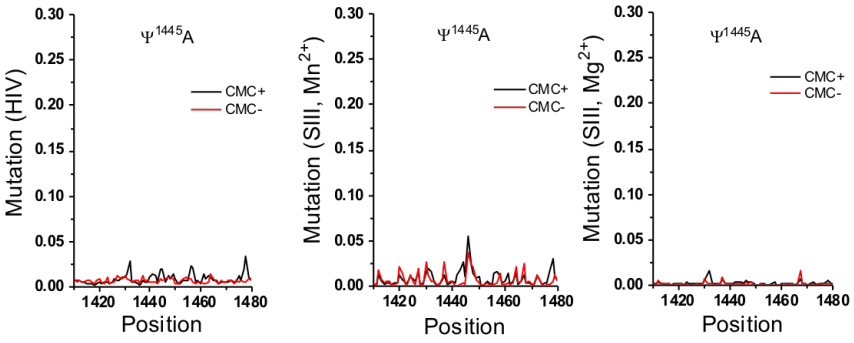


Figure 5.3: Comparison of libraries prepared with different RT enzymes. Box-and-whisker plots of Δmutation and Δstop rates at known Ψ sites in rRNA, in RNA libraries prepared with HIV RT, SuperScript III RT with Mn^{2+}, or SuperScript III RT with Mg^{2+}. ΔRate = rate in CMC-treated sample − rate in untreated sample.



(a)



(b)

Figure 5.4: Stop and mutation rates around selected Ψ sites. (a) Stop rates around Ψ1004 in18S rRNA, which has a +1 G context, in libraries prepared with HIV RT, SuperScript III RT with Mn^{2+}, or SuperScript III RT with Mg^{2+}. (b) Mutation rates around Ψ1445 in 18S rRNA, which has a +1 A context, in libraries prepared with HIV RT, SuperScript III RT with Mn^{2+}, or SuperScript III RT with Mg^{2+}.

CMC-Ψ read-through led to skipped or added nucleotides, we re-aligned our sequencing data to allow deletions and insertions. Deletion rates were < 0.03 at all but 4 out of 36 pseudouridine sites in the 18S rRNA, whereas insertion rates occurred at similarly low levels in the libraries prepared from CMC-treated and mock-treated rRNA. Therefore, we did not account for deletions or insertions in our subsequent analysis. While lower than expected, Δmutation rates in the libraries prepared using HIV RT or SIII RT, Mn^{2+} were sufficient to identify certain pseudouridine sites, such as Ψ1004 in 18S rRNA, which has a G at the +1 position, i.e. one nucleotide $3'$ to the pseudouridine (Figures 5.2(b) and 5.4(a)). In contrast,

other pseudouridine sites, such as Ψ1445 in 18S rRNA, which has an A at the +1 position, were best identified based on Δstop rates in libraries prepared using any of the three reverse transcription conditions (Figures 5.2(c) and 5.4(b)).

5.2.3 Context-dependent mutation and stop rates at pseudouridine sites

We examined the effect of sequence context on mutation and stop rates at pseudouridine sites in CMC-treated and mock-treated rRNA. We limited this analysis to the sequencing libraries prepared with HIV RT, which had higher median Δmutation rates at known pseudouridine sites than libraries prepared with SIII RT, Mn^{2+} or SIII RT, Mg^{2+} (Figure 5.3). We found that the nucleotide $3'$ to pseudouridine in the RNA template (the +1 nucleotide) influenced the Δmutation and Δstop rates at known pseudouridine sites, but not at unmodified uridine sites, in human 18S and 28S rRNAs (Figure 5.5(a)–5.5(c)). Known pseudouridines with G at the +1 position (+1 G) tended to have high Δmutation rates and low Δstop rates, whereas known pseudouridines with A at the +1 position (+1 A) tended to have high Δstop rates and low Δmutation rates (Figures 5.5(a) and 5.5(c)). The Δmutation and Δstop rates at known pseudouridines with C or U at the +1 position (+1 C or +1 U) tended to follow the same patterns as the rates at known pseudouridines with +1 G or +1 A, respectively, but these patterns were less pronounced (Figure 5.5(a)). Overall, known pseudouridines with +1 C or +1 U could have either elevated Δmutation rates or elevated Δstop rates (Figure 5.5(a)), resulting in intermediate median Δmutation and Δstop rates (Figure 5.5(c)). The trend in context-dependent Δmutation rates (+1 G > +1 U ∼ +1 C > +1 A) was not fully explained by the trend in Δstop rates, since this pattern remained when stops were not counted (median Δmutation rates without counting stops: 0.104 for +1 G, 0.056 for +1 U, 0.037 for +1 C, and 0.023 for +1 A). Next, we examined the impact of the +1 nucleotide on the identity of the bases that were misincorporated opposite of CMC-Ψ during reverse transcription, or the mutation signature (Figure 5.5(d)). Regardless of the +1 nucleotide, known pseudouridine sites in rRNA had their highest Δmutation rate to C, and a higher

Δmutation rate to A than to G, indicating that G tended to be misincorporated most often, and U was misincorporated more often than C. Thus, the total Δmutation and Δstop rates at

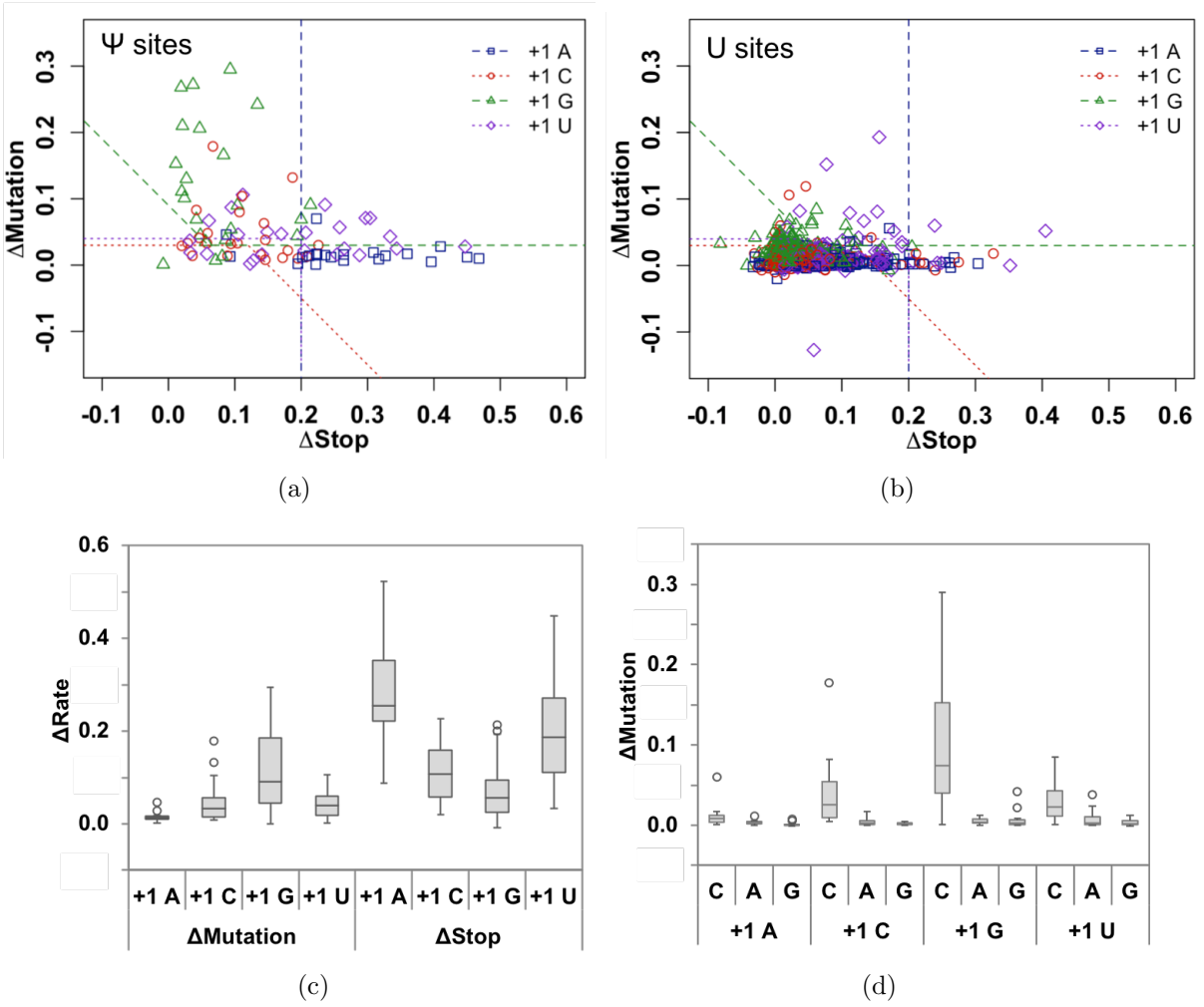


Figure 5.5: Context-dependent mutation and stop rates at Ψ sites. ΔMutation and Δstop rates at all known Ψ sites (a) and at all unmodified U's (b) in 18S and 28S rRNA, in RNA libraries prepared with HIV RT. Dashed and dotted lines represent the context-dependent thresholds used to identify Ψ sites (+1 A: Δstop > 0.2; +1 C: ΔMI > 0.15 OR Δmutation > 0.03; +1 U: Δstop > 0.2 OR Δmutation > 0.04; +1 G: ΔMI > 0.09 AND Δmutation > 0.03). Color and shape specifies the +1 nucleotide context. Mismatch index (MI) = mutation rate + stop rate. ΔRate = rate in CMC-treated sample (CMC+) − rate in mock-treated sample (CMC−). (c) Context-dependent mutation and stop rates. Box-and-whisker plots of Δmutation and Δstop rates at known Ψ sites with different +1 nucleotide contexts, in libraries prepared with HIV RT. (d) Context-dependent mutation signatures. Box-and-whisker plots of Δmutation rate to C, A, or G at known Ψ sites with different +1 nucleotide contexts, in libraries prepared with HIV RT.

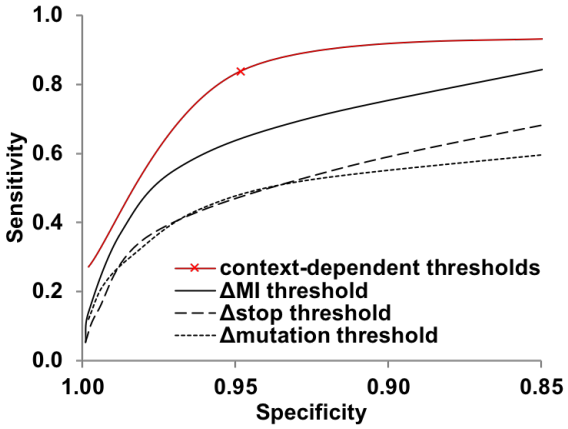


Figure 5.6: Receiver operating characteristic (ROC) curves for Ψ site detection in 18S and 28S rRNA. For context-dependent thresholds, Δmutation, Δstop, and/or Δmismatch index (MI) thresholds vary depending on the +1 nucleotide. The thresholds used to construct this curve were selected to maximize sensitivity and specificity. The red X marks the context-dependent thresholds shown in Figure 5.5(a)–5.5(b). Context-dependent thresholds can increase both sensitivity (shift up) and specificity (shift left) of Ψ site detection, as compared to Δmutation, Δstop, or ΔMI thresholds that do not account for +1 nucleotide context.

known pseudouridine sites in rRNA varied depending on the +1 nucleotide, but the mutation signature was not strongly dependent on the +1 nucleotide.

Previous pseudouridine sequencing studies have relied on stop rates or other stop-based metrics to set thresholds for calling pseudouridine sites [66–68, 179, 184]. Using our rRNA sequencing data, we compared the performance of Δstop rate thresholds to the performance of Δmutation rate thresholds and of thresholds in a combined metric, Δmismatch index (MI = stop rate + mutation rate). Since Δmutation and Δstop rates were dependent on +1 nucleotide context, we also tested context-dependent thresholds, in which different Δmutation and Δstop rate thresholds were established for different +1 nucleotide contexts. We found that Δmutation rate thresholds and Δstop rate thresholds performed similarly, Δmismatch index thresholds performed better, and context-dependent thresholds had the highest sensitivity and specificity for pseudouridine sites in human rRNA (Figure 5.6).

5.3 Discussion

In this study, we identified reverse transcription conditions that favor CMC-Ψ read-through, and we showed that pseudouridines in CMC-treated rRNA have context-dependent mutation and stop rates that can be used to improve pseudouridine site detection in high-throughput sequencing. Using either a different reverse transcriptase enzyme, HIV RT, or a different divalent cation with SuperScript III RT, Mn^{2+}, led to over 80% read-through of CMC-Ψ in a synthetic RNA oligo (Figure 5.1). These reverse transcription conditions also decreased the stop rate and increased the mutation rate at pseudouridine sites in a CMC-treated human rRNA sequencing library (Figure 5.3), though not to the same degree as expected based on the efficient CMC-Ψ read-through observed with the model oligo. This discrepancy could be due to changes in reverse transcriptase efficiency and processivity in the setting of extracted cellular RNA, as the reverse transcription reaction is highly sensitive to reaction parameters including RNA concentration [286, 287]. The low mutation rate at pseudouridine sites in CMC-treated rRNA could also be due to the incorporation of A opposite of CMC-Ψ, which would produce neither a stop nor a mutation. In fact, the low Δmutation rate and the mutation signature observed at known pseudouridine sites in rRNA (Figures 5.3 and 5.5(d)) follow the same patterns as non-templated synthesis, in which HIV RT prefers to add A or G, with a lower preference for T [288], suggesting that HIV RT conducts non-templated synthesis upon encountering CMC-Ψ.

Although the stop rate was higher and the mutation rate lower than expected based on our results with the model oligo, these sequencing results allowed us to incorporate both stop and mutation information in our analysis. Consistent with previous results with a different RNA modification [188], the combined analysis of stop and mutation rates enhanced the prediction of modification sites relative to evaluating stops or mutations alone (Figure 5.6). Previous work has also shown that the nucleotide $3'$ to an RNA modification site can affect its mutation signature [188]. In our sequencing data, the +1 nucleotide did not influence the mutation signature at pseudouridine sites in CMC-treated rRNA, but it did significantly

impact the stop and mutation rates. Known pseudouridines with +1 G or, to a lesser extent, +1 C, tended to have higher Δmutation rates and lower Δstop rates than average, whereas known pseudouridines with +1 A or, to a lesser extent, +1 U, tended to have higher Δstop rates and lower Δmutation rates than average (Figure 5.5(a)–5.5(c)). These patterns can be explained by the higher stability of G-ribonucleotide–C-deoxyribonucleotide (rG–dC) and rC–dG base pairs relative to rA–dT and rU–dA base pairs [289], since a more stable base pair between the +1 nucleotide and the 3′ end of the cDNA could prevent fraying of the RNA–cDNA hybrid and thereby increase the efficiency of cDNA extension when the reverse transcriptase encounters CMC-Ψ. The dependence of the measured stop rates on +1 nucleotide context could also result from 3′-end nucleotide biases in the circularization step. The observed stop rates follow the pattern +1 A > +1 U > +1 C > +1 G. Reverse transcription stops at pseudouridine sites with these +1 positions in the template RNA would lead to truncated cDNAs with 3′-end nucleotides T, A, G, and C, respectively. Therefore, the observed pattern in stop rates matches the 3′-end nucleotide preferences of CircLigase I (3′ T > 3′ A > 3′ G, with no detectable ligation for 3′ C, according to Epicentre). Although the end-base preferences of CircLigase II have not been reported, CircLigase II is an adenylated form of CircLigase I and therefore mostly likely has similar sequence preferences. However, the trend in mutation rates (+1 G > +1 U ∼ +1 C > +1 A) was observed even when stops were not counted, which cannot be explained by differences in either RNA–cDNA hybrid stability or circularization efficiency.

Based on our observations, we created simple and interpretable context-dependent thresholds for the identification of pseudouridine sites based on high-throughput sequencing of CMC-treated and mock-treated RNA. These context-dependent thresholds required that putative pseudouridines with +1 G have high Δmutation rates, putative pseudouridines with +1 A have high Δstop rates, and putative pseudouridines with +1 C or +1 U have either high Δmutation rates or high Δstop rates, in order to be identified as pseudouridine sites (Figure 5.5(a)). These context-dependent thresholds improved the detection of pseu-

douridine sites relative to context-independent thresholds (Figure 5.6). Given the sensitivity of the reverse transcription reaction to various reaction parameters [286, 287], the trends in mutation and stop rates at pseudouridine sites in CMC-treated human rRNA, which is highly structured and modified, might differ from the patterns in other RNAs such as mRNA. Nonetheless, the context-dependent patterns observed in rRNA provide a reasonable starting point for the detection of pseudouridines in high-throughput sequencing of other CMC-treated RNA samples. Moreover, regardless of the specific mutation and stop rates, accounting for context-dependence of mutation and stop rates by using context-dependent thresholds could increase the sensitivity and specificity of pseudouridine detection. Thus, context-dependent mutation and stop rates provide valuable information that may enhance the detection of pseudouridine or other RNA modifications in high-throughput sequencing.

5.4 Materials and Methods

5.4.1 RNA oligo synthesis and CMC modification

Oligonucleotide synthesis was performed on an Expedite Nucleic Acid Synthesis System using standard RNA synthesis conditions on a 1-μmol scale. Controlled Pore Glass (CPG) supports and A, C, G, and T phosphoramidites were purchased from ChemGene. Pseudouridine phosphoramidite was purchased from Glen Research. The terminal 4,4′-dimethoxytrityl (DMTr) protecting group was removed from the oligonucleotides by using the DMTr-off mode. The resin containing the oligos was transferred to a 2-mL vial, and a mixture of 0.3 mL ethanol and 0.9 mL 30% ammonium hydroxide was added. The vial was then incubated at 55 °C for 4 hours to remove all the base-labile protecting groups. Once cooled to room temperature, the supernatant was transferred to a 1.5-mL tube and dried in a SpeedVac centrifugal vacuum concentrator. A mixture of 100 μL dimethyl sulfoxide (DMSO) and 125 μL hydrogen fluoride triethylamine (Sigma Aldrich) was added to the tube, and the tube was incubated in a 65 °C water bath for 2.5 hours. After cooling to room temperature, 22.5

μL 3 M sodium acetate (pH 5.3) was added, and the mixture was vortexed. Next, 1 mL of n-butanol was added, and the mixture was kept at -80 °C overnight. After spinning at maximum speed for 25 minutes at 4 °C, the pellet was washed with 1 mL of 70% v/v ethanol and spun again. The pellet was then dissolved in 1 mL of nuclease-free water and purified by C18 reverse-phase high-performance liquid chromatography (HPLC) with the applied buffer as 0–20% acetonitrile in 0.1 M triethylammonium acetate in water. The fractions were checked for purity by analytical HPLC and matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS). The purified oligonucleotides were concentrated in a SpeedVac centrifugal vacuum concentrator. The sequences of the synthesized RNA oligos were as follows:

Oligo Ψa: 5′-UACACUCAGΨUCGGACUAAAGCUGCUC

Oligo Ua: 5′-UACACUCAGUUCGGACUAAAGCUGCUC

To prepare the CMC-Ψ-containing RNA oligo, 10 μg of Oligo Ψa in 12 μL water were combined with 24 μL of 50 mM Tris-Cl (pH 8.3), 4 mM EDTA, 7 M urea (TEU buffer) and 4 μL of 1 M *N*-cyclohexyl *N*′-(2-morpholinoethyl) carbodiimide (CMC) freshly dissolved in TEU buffer, for a final concentration of 0.1 M CMC. The mixture was incubated at 30 °C overnight (16 hours), and then the RNA was purified with an Oligo Clean & Concentrator column (Zymo Research, D4061) and eluted in 20 μL water. Next, 40 μL of 50 mM sodium carbonate and 2 mM EDTA (pH 10.4) were added, and the 60-μL reaction was incubated at 45 °C for 2 hours or at 37 °C for 4 hours. Analytical HPLC and MALDI-MS showed that almost all CMC on U's and G's were removed, while ∼80% of CMC-Ψ adducts remained.

5.4.2 Purification, radiolabeling, and phosphorylation of DNA primers and DNA ladder

The following DNA primers were purchased from Integrated DNA Technologies (IDT):

Primer A: 5′-GAGCAGCTTTAG

Primer B: 5′-GATCGTCGGACTGTAGAACTAGACGTGTGCTCTTCCGATCT

Primer 18S-831: 5′-GTATCCAGGCGGCTCGGGCC

Primer 28S-3755: 5′-GATGACGAGGCATTTGGCTACC

Illumina multiplex primer:

5′-AATGATACGGCGACCACCGAGATCTACACGTTCAGAGTTCTACAGTCCGACGATC

Barcoded primer: 5′-CAAGCAGAAGACGGCATACGAGAT[Barcode]GTGACTGGAGTTCA GACGTGTGCTCTTCCGATCT

Primers A, B, 18S-831, and 28S-3755, and the Illumina multiplex and Barcoded primers, were purified on a gel containing 10% acrylamide:bisacrylamide (29:1), 7 M urea, 89 mM Tris-borate (pH 8.3), and 2 mM Na_2EDTA (ethylenediaminetetraacetic acid). DNA was excised from the gel by UV shadowing and eluted in 50 mM potassium acetate and 200 mM KCl (pH 7.5) by the crush-and-soak method. Eluted RNA was precipitated in ethanol, and then resuspended and stored in water at −20 °C. Primers A, 18S-831, and 28S-3755 were 5′ ^{32}P-labeled with 20 pmol of 6000 Ci/mmol γ-^{32}P-ATP per 100 pmol of primer, using 10 U of T4 polynucleotide kinase (T4 PNK, New England BioLabs, M0201L) in 70 mM Tris-Cl (pH 7.6), 10 mM $MgCl_2$, and 5 mM DTT (1× PNK buffer) in a total volume of 10 μL at 37 °C for 30–60 minutes, and then extracted once with phenol:chloroform (3:1), ethanol precipitated, and resuspended to a concentration of 2 μM RNA in water. Primer B was 5′ ^{32}P-labeled with 13.2 pmol of 6000 Ci/mmol γ-^{32}P-ATP per 300 pmol of primer, using 10 U of T4 PNK in 1× PNK buffer in a total volume of 20 μL at 37 °C for 20–30 minutes. Next, 1.2 nmol (4 molar equivalents) of cold ATP and an additional 10 U of T4 PNK were added, and the reaction was incubated at 37 °C for 15 minutes to fully phosphorylate the primer. Following one round of phenol:chloroform (3:1) extraction, the primer was ethanol precipitated and resuspend to a final concentration of 30 μM in water.

The 10 bp DNA ladder (Invitrogen, 10821-015) was labeled by PNK exchange: 0.4 μg of 10 bp ladder was combined with 50 pmol of ATP and 20 U of T4 PNK in 50 mM imidazole pH 6.4, 12 mM $MgCl_2$, 1 mM 2-mercaptoethanol, and 70 μM ADP in a total volume of 10 μL and incubated at 37 °C for 40 minutes, and then extracted once with phenol:chloroform

(3:1), ethanol precipitated, and resuspended in 9 M urea, 100 mM EDTA, 0.2% w/v xylene cyanol, and 0.2% w/v bromophenol blue.

5.4.3 Adapter adenylation

The 3′ DNA adapter was purchased from IDT with a 5′ phosphate and a 3′ inverted dT blocking group: /5Phos/AGATCGGAAGAGCACACGTCTAGTTCTACAGTCCGACGATC/3invdT/. Adenosine 5′-phosphorimidazolide (ImpA) was synthesized according to published protocols [290, 291]. The 3′ DNA adapter was 5′-adenylated in 50 mM freshly dissolved ImpA and 25 mM $MgCl_2$ at 50 °C for 3 hours, and then purified on a gel containing 20% acrylamide:bisacrylamide (29:1), 7 M urea, 89 mM Tris-borate (pH 8.3), and 2 mM Na_2EDTA by the crush-and-soak method.

5.4.4 Reverse transcription of RNA oligos

For annealing, 1 pmol of Primer A was mixed with 0.8–1 pmol of CMC-modified or unmodified Oligo Ψa in 50 mM Tris-Cl (pH 7.4) and 100 mM KCl in a total volume of 2.5–5 μL. The primer–template mixture was incubated at 93 °C for 2 minutes, and then on ice for 4 minutes. Reverse transcription with either 0.6 U/μL AMV RT (New England BioLabs, M0277T) or 0.28 U/μL HIV RT (Worthington Biochemical Corporation, LS05003) was conducted in 50 mM Tris-acetate (pH 8.3), 75 mM potassium acetate, 8 mM magnesium acetate, 10 mM DTT, and 0.5 mM of each dNTP in a total volume of 5–10 μL. Reverse transcription with SuperScript III RT (Thermo Scientific, 18080044) was conducted in 50 mM Tris-Cl (pH 8.3), 75 mM KCl, 10 mM DTT, and the specified concentration of $MgCl_2$ or $MnCl_2$ in a total volume of 5–10 μL. The reactions were incubated at 42 °C for 15 minutes for AMV RT, at 37 °C for 1 hour for HIV RT, or at 42 °C for 3 hours for SuperScript III RT. Next, 10 μL of 9 M urea, 100 mM EDTA, 0.2% w/v xylene cyanol, and 0.2% w/v bromophenol blue were added, and the samples were incubated at 90 °C for 5–10 minutes before loading on a gel containing 20% acrylamide:bisacrylamide (29:1), 7 M urea, 89 mM Tris-borate (pH

8.3), and 2 mM Na_2EDTA. After gel electrophoresis, primer extension was visualized by phosphorimaging.

5.4.5 Preparation of sequencing libraries

Human embryonic kidney (HEK) cell line HEK293T/17 was obtained from the American Type Culture Collection (ATCC) and cultured in Dulbecco's Modified Eagle's Medium (DMEM) with high glucose and L-glutamine, without sodium pyruvate (HyClone, SH30022.01) in a 37 °C incubator with a humidified atmosphere of 5% CO_2. Total RNA was isolated from confluent 100-mm plates of HEK293T cells by Trizol extraction (Life Technologies, 15596-018) and size-selected for RNAs over 200 nucleotides long with the mirVana miRNA isolation kit (Life Technologies, AM1561). The integrity of size-selected total RNA was checked on a gel containing 1.2% w/v agarose, 89 mM Tris-borate (pH 8.3), 2 mM Na_2EDTA, and 0.4 μg/mL ethidium bromide. The RNA was diluted to 125 ng/μL and fragmented in 50 mM Tris-Cl (pH 7.9) and 8 mM $MgCl_2$ at 94°C for 10 min, and then purified with an Oligo Clean & Concentrator column (Zymo Research), yielding a total of 34 μg of fragmented RNA. The size distribution of fragmented RNA was found to be centered around 200 nucleotides on a gel containing 1.5% w/v agarose, 89 mM Tris-borate (pH 8.3), 2 mM Na_2EDTA, and 0.4 μg/mL ethidium bromide. The 50%-Ψ standard was added at 0.06 pmol per 1 μg of fragmented RNA, corresponding to 0.03 pmol/μg each of Oligos Ψa and Ua. The RNA was separated into three parts with 11.3 μg RNA each: two parts for CMC treatment (CMC+) and one part for mock treatment (CMC−). The RNA in all three parts was denatured in a total volume of 10 μL at 80 °C for 2 minutes, and then placed on ice. The denatured RNA was combined with 20 μL of 50 mM Tris-Cl (pH 8.3), 4 mM EDTA, 7 M urea (TEU buffer) and either 20 μL of TEU buffer (for CMC−) or 20 μL of 1 M CMC freshly dissolved in TEU buffer for a final concentration of 0.4 M CMC (for CMC+). The reactions were incubated at 40 °C for 30 minutes, after which the RNA was purified with an Oligo Clean & Concentrator column (Zymo Research) and eluted with 20 μL of water. Next, 40 μL of 50 mM sodium

carbonate and 2 mM EDTA (pH 10.4) were added, and the reaction was incubated at 45 °C for 2 hours. The samples were ethanol precipitated and resuspended in 20 μL of water.

The RNA was combined with 1 U of calf intestinal alkaline phosphatase (New England BioLabs, M0290S) and 40 U of T4 PNK in 1× PNK buffer in a total volume of 30 μL, and the dephosphorylation reaction was incubated at 37 °C for 30 minutes. After extraction with phenol:chloroform (3:1), the RNA was ethanol precipitated and resuspended in 20 μL water. The 5′-adenylated 3′ DNA adapter was ligated to the repaired RNA at an estimated 1:2 molar ratio of adaptor to RNA (24.25 pmol adaptor for ∼48.5 pmol or 3.3 μg of 200-nucleotide-long fragmented RNA) with 10 U/μL of T4 RNA ligase 2, truncated KQ (New England BioLabs) in 50 mM Tris-Cl (pH 7.5), 10 mM $MgCl_2$, 1 mM DTT, 15% v/v PEG8000 in a total volume of 40 μL at 16 °C overnight (∼16 hours). The RNA was purified with an Oligo Clean & Concentrator column (Zymo Research) and eluted in 14 μL of water. The 5′ ^{32}P-labeled Primer B (36.3 pmol each for CMC+ and CMC−) was annealed to the RNA in 50 mM Tris-Cl (pH 7.4) and 100 mM KCl in a total volume of 16.5 μL at 93 °C for 2 minutes, and then placed on ice for 3 minutes. The CMC+ and CMC− RNA samples were each split into three parts for reverse transcription, where each part contained 1 μg RNA and 11 pmol 5′ ^{32}P-labeled Primer B. Reverse transcription was conducted in 10-μL reactions, with (1) HIV RT: 0.28 U/μL HIV RT, 50 mM Tris-Cl (pH 8.3), 75 mM potassium acetate, 8 mM magnesium acetate, 10 mM DTT, and 0.5 mM of each dNTP at 37 °C for 1 hour; (2) SIII RT, Mn^{2+}: 10 U/μL SuperScript III RT, 50 mM Tris-Cl (pH 8.3), 75 mM KCl, 3 mM MnCl2, 10 mM DTT, and 0.5 mM of each dNTP at 42 °C for 3 hours; or (3) SIII RT, Mg^{2+}: 10 U/μL Superscript III RT, 50 mM Tris-Cl (pH 8.3), 75 mM KCl, 8 mM $MgCl_2$, 10 mM DTT, and 0.5 mM of each dNTP at 42 °C for 3 hours. To end the reverse transcription reaction, 10 μL of 9 M urea, 100 mM EDTA, 0.2% w/v xylene cyanol, and 0.2% w/v bromophenol blue were added, and sample was incubated at 90 °C for 5 minutes. The cDNA was separated on a gel containing 7.5% acrylamide:bisacrylamide (29:1), 7 M urea, 89 mM Tris-borate (pH 8.3), and 2 mM Na_2EDTA, along with 5′ ^{32}P-labeled 10 bp DNA ladder (Invitrogen, 10821-015)

and 5′ ^{32}P-labeled Primer B as size controls. After gel electrophoresis, primer extension was visualized by phosphorimaging. The 55-to-260-nucleotide region of each lane was excised, and the cDNA was eluted in 50 mM potassium acetate and 200 mM KCl (pH 7.5) by the crush-and-soak method, and then ethanol precipitated and resuspended in water.

The cDNA was circularized with 5 U/μL of CircLigase II (Epicentre, CL4115K) in 33 μM Tris-acetate (pH 7.5), 66 μM potassium acetate, 0.5 mM DTT, 2.5 mM MnCl2, and 1 M betaine at 60 °C overnight, and then the ligase was inactivated at 80 °C for 10 minutes. The cDNA was extracted with phenol-chloroform (3:1) and with chloroform, and then ethanol precipitated and resuspended in water. The cDNA was amplified with 200 μM Illumina multiplex primer and 200 μM Barcoded primer in 1× Phusion High-Fidelity PCR Master Mix with HF buffer (Thermo Scientific, F531) in a total volume of 50 μL: 30 seconds at 98 °C; 12x: 10 seconds at 98 °C, 30 seconds at 60 °C, 30 seconds at 72 °C; 5 minutes at 72 °C; hold at 4 °C. The cDNA was extracted with phenol:chloroform (3:1) and with chloroform, and then ethanol precipitated. The cDNA was then purified on a 6% TBE minigel (Novex, EC6265BOX) with a 100-base-pair DNA ladder with ethidium bromide staining by the crush-and-soak method, ethanol precipitated, and resuspended in 15 μL of water. The library was checked by Bioanalyzer capillary electrophoresis and by quantitative PCR, and then sequenced in a single lane by Illumina HiSeq 4000 with paired-end 100-base-pair reads at the University of Chicago Genomics Facility.

5.4.6 Mapping of sequencing data

Standard quality control using FastQC was performed after the sequencing and trimming steps. The sequencing reads were processed first with Trimmomatic v0.32 and then with custom Python scripts designed to remove any additional artifacts from demultiplexing and removal of primers, adapters, and low-quality sequences [292]. Next, the trimmed sequences were simultaneously aligned to the 18S and 28S rRNA reference sequences (NR_003286 and NR_003287.2) using Bowtie 1.0 with the highest allowed mismatch settings, yielding 20–50

million 18S and 28S rRNA aligned reads per sample.

The sequence alignment/map (SAM) output from Bowtie was further processed using Python scripts, first to remove a single nucleotide from either end of each read (end-trimming), and then to determine the total read count (c), the number of mutations (m), and the number of stops (s) at each position of the 18S and 28S rRNA reference sequences [292]. Stops at nucleotide position n in the rRNA sequence correspond to reads with a 5′-most nucleotide aligning to the $n + 2$ position, which accounts for end-trimming. The stop rate at each position in the human 18S and 28S rRNAs was calculated as $s/(c + s)$. The mutation rate at each position in the human 18S and 28S rRNAs was calculated as $m/(c+s)$. The mutation rate without counting stops was calculated as m/c. For total mutation rate, m was the sum of A, C, and G reads; for the mutation rate to A, C, or G, m was the number of A, C, or G reads. The mismatch index (MI) at each position was calculated as stop rate + mutation rate. The Δstop and Δmutation rates, and Δmismatch index, were calculated as: *rate or index in the CMC-treated sample (CMC+) − rate or index in mock-treated sample (CMC−).*

To determine deletion rates, the trimmed sequences were aligned to the 18S and 28S rRNA reference sequences using Bowtie 2.0, and then visualized using the Integrative Genomics Viewer. At each position in the reference sequences, the number of deletions (d) and the total read count (c) at that position were used to calculate the deletion rate as d/c.

5.4.7 Identification of pseudouridine sites

A list of modified sites in human 18S and 28S rRNAs was obtained from the Fournier lab's 3D rRNA modification maps database [285], and the nucleotide positions were re-assigned according to the National Center for Biotechnology Information sequences for human 18S rRNA (NR_003286) and 28S rRNA (NR_003287.2). Having excluded 2′-*O*-methylated pseudouridine, 2′-*O*-methylated uridine, and 1-methyl-3-(3-amino-3-carboxypropyl)pseudouridine from the analysis, we counted 92 known pseudouridines and 1 047 unmodified uridines in the hu-

man 18S and 28S rRNAs.

The context-dependent thresholds for pseudouridine identification were chosen by separately evaluating possible thresholds for each possible +1 nucleotide context. For +1 A, Δstop rate and ΔMI thresholds were evaluated. For +1 C and +1 G, combinations of ΔMI and Δmutation rate thresholds were evaluated. For +1 U, combinations of Δstop rate, Δmutation rate, and ΔMI thresholds were evaluated. In each case, receiver operating characteristic (ROC) curves were plotted, and the threshold with the highest performance in sensitivity, specificity, positive predictive value, and negative predictive value was chosen. The optimized thresholds for all four +1 nucleotide contexts were combined to define the context-dependent threshold shown in Figures 5.5(a)–5.5(b) (+1 A: Δstop $>$ 0.2; +1 C: ΔMI $>$ 0.15 OR Δmutation $>$ 0.03; +1 U: Δstop $>$ 0.2 OR Δmutation $>$ 0.04; +1 G: ΔMI $>$ 0.09 AND Δmutation $>$ 0.03).

The ROC curve for ΔMI threshold was formed by varying the ΔMI threshold from 0.1 to 0.45; the ROC curve for Δstop threshold was formed by varying the Δstop rate threshold from 0.08 to 0.4; the ROC curve for Δmutation threshold was formed by varying the Δmutation rate threshold from 0.02 to 0.12. The ROC curve for context-dependent thresholds was formed by simultaneously varying the ΔMI, Δstop rate, and/or Δmutation rate thresholds for all four +1 nucleotide contexts: for +1 A, the Δstop rate threshold was varied from 0.08 to 0.4; for +1 C, the ΔMI threshold was varied from 0.15 to 0.35, and the Δmutation rate threshold from 0.025 to 0.03; for +1 G, the ΔMI threshold was varied from 0.07 to 0.15, and the Δmutation rate threshold from 0.012 to 0.08; for +1 U, the Δstop rate threshold was varied from 0.2 to 0.25, and the Δmutation rate threshold from 0.015 to 0.04. Context-dependent thresholds within these ranges were compared, and the thresholds with the high sensitivity and specificity were used to form the ROC curve shown in Figure 5.6.

5.4.8 Data Deposition

The sequencing data have been deposited to National Center for Biotechnology Information Gene Expression Omnibus database under accession number GSE110247.

Chapter 6

Conclusions and Future Directions

Gene expression is controlled by a complex and interconnected regulatory network. Through their impact on both RNA structure and protein binding, RNA modifications can regulate every step in the post-transcriptional lifespan of an RNA transcript. Furthermore, RNA modifications that occur co-transcriptionally can both influence and be influenced by the transcriptional process. In this way, RNA modifications illustrate the diversity and interdependence of the mechanisms used by the cell to regulate gene expression.

In this thesis, I have explored relationships between RNA modifications, RNA structure, RNA-binding proteins, and the transcriptional machinery. First, I examined how the abundant mRNA modification m^6A alters the structure of an RNA hairpin and facilitates binding of the protein hnRNPC. Second, we discovered a novel m^6A reader protein, hnRNPG, which uses a similar mechanism to selectively bind m^6A-modified RNA, but is distinct in that it binds methylated transcripts by using RGG motifs in its low-complexity region. Third, I studied the cellular functions of hnRNPG, finding that it interacts directly with the phosphorylated CTD of RNAPII in addition to selectively binding m^6A-modified RNAs. I proposed that co-transcriptional binding of hnRNPG, both to the CTD of transcribing RNAPII and to m^6A-modified nascent RNA, is required for the m^6A-dependent regulation of alternative splicing by hnRNPG. Finally, I attempted to improve on current pseudouridine sequencing methods by using mutations in addition to stops to identify pseudouridine sites, and I found that CMC-modified pseudouridines produce context-dependent mutations and stops in high-throughput sequencing of human rRNA. Together, these studies illustrate how RNA modifications modulate RNA structure, protein binding and function, and sequencing signatures. At the same time, this work raises questions regarding the effect of RNA modifications on RNA structure and protein binding, possible mechanisms for co-transcriptional gene regulation by RNA modifications and their reader proteins, and the broader applicability of combined mutation- and stop-based RNA modification sequencing methods.

6.1 Structure-Dependent Binding to Modified RNAs

Although m^{6}A does not preclude Watson–Crick base pairing, the N^6-methyl group alters the stability of RNA secondary structure. These m^{6}A-induced changes in RNA structure can facilitate the binding of Class II m^{6}A reader proteins, which use an m^{6}A-switch mechanism to selectively bind modified RNAs. The first Class II m^{6}A reader protein to be discovered was hnRNPC, which was found to use an m^{6}A-switch mechanism to selectively bind an m^{6}A-modified hairpin in the lncRNA *MALAT1* [119]. In Chapter 2 of this thesis, I further characterized this first example of an m^{6}A switch in a cellular RNA [1]. I used NMR and FRET to demonstrate the effect of m^{6}A on an RNA hairpin derived from the m^{6}A switch in *MALAT1*. The observed imino proton NMR resonances and FRET efficiencies suggested that m^{6}A selectively destabilizes the portion of the hairpin-stem where the hnRNPC binding site is located, increasing the solvent accessibility of the neighboring bases while maintaining the overall hairpin structure. The m^{6}A-modified hairpin has a predisposed conformation that resembles the hairpin conformation in the RNA–hnRNPC complex more closely than the unmodified hairpin. The m^{6}A-induced structural changes in the *MALAT1* hairpin serve as a model for other m^{6}A switches that likely affect the binding of hnRNPC and other Class II m^{6}A reader proteins.

Notably, the effect of m^{6}A modification on the structure of the *MALAT1* hairpin is subtle. Rather than completely destabilizing the RNA duplex to yield single-stranded RNA, m^{6}A only locally destabilizes a portion of the hairpin to promote hnRNPC binding. hnRNPC is a single-stranded RNA-binding protein, yet the m^{6}A-induced destabilization of the hairpin is sufficient to favor hnRNPC binding to the methylated *MALAT1* hairpin by ∼8-fold relative to the nonmethylated *MALAT1* hairpin. The subtle effect of m^{6}A on RNA structure raises the possibility that some m^{6}A reader proteins might recognize features such as bulges or loops in RNA duplexes, rather than binding to single-stranded RNA. In fact, this type of specificity could apply in the case of the Class II m^{6}A reader hnRNPG, which I investigated in Chapters 3–4 [3]. hnRNPG selectively binds to m^{6}A-modified RNA through

a low-complexity region that includes RGG repeats. RGG repeats tend to have degenerate specificity for RNA sequences and structures [210]. However, several RGG regions have been shown to preferentially bind RNA helices perturbed by local features, since such perturbations facilitate insertion of the RGG repeats into the major groove of the imperfect A-form RNA helix [208–210]. It is therefore conceivable that m^6A modification might similarly perturb structured RNAs and thereby promote binding by an RGG region of hnRNPG. This mechanism would be a variation on the m^6A-switch mechanism, in which m^6A creates a structural feature – namely, a perturbed A-form helix – that is recognized by an m^6A reader protein, rather than exposing a sequence that is selectively bound by a single-stranded RNA-binding protein. The structure-specificity of hnRNPG could be investigated in more detail by generating structural variants of the *MALAT1* hairpin and evaluating their ability to pull down hnRNPG or cross-link its C-terminal RGG region. Another approach to this question would be to investigate in the structures of the hnRNPG-bound m^6A sites identified by PAR-CLIP–MeRIP, both experimentally and through structure prediction. Recurrent patterns in the structures of these hnRNPG-bound m^6A sites could identify m^6A-induced structural features that are specifically recognized by hnRNPG.

The m^6A modification of RNA can also have other effects on RNA structures, beyond the destabilization of RNA duplexes. The effect of m^6A on RNA duplex structures is position-dependent. While m^6A modifications within a duplex are destabilizing, m^6A at the end of a duplex increases stacking interactions and consequently stabilizes RNA duplexes by 0.42–0.58 kcal/mol [139]. Therefore, m^6A reader proteins that recognize an m^6A-induced change in RNA structure could potentially be selective for RNA duplexes over single-stranded RNA or perturbed RNA helices. Moreover, the N^6-methyl can occur on either the Hoogsteen or Watson–Crick edge of adenosine, depending on whether it adopts an *anti* or *syn* conformation, respectively. Thus, m^6A could potentially disrupt interactions involving the Hoogsteen edge, such as Hoogsteen base pairs or base triples [70]. The disruption of such interactions could expose RNA sequence or structure motifs bound by m^6A reader proteins. Alterna-

tively, the effects of m^6A on the secondary and tertiary structure of RNA could decrease binding by 'anti-reader' RNA-binding proteins. Several m^6A 'anti-readers' have been discovered, including the human single-stranded RNA-binding protein Pumilio 2 (hPUM2) and the stress granule proteins Ras GTPase-activating protein-binding proteins 1 and 2 (G3BP1 and G3BP2) [175–177]. However, it is unclear whether the effect of m^6A on these RNA–protein interactions results from the interference of m^6A with direct RNA–protein contacts or from an m^6A-induced change in RNA structure. G3BP1 and G3BP2 binding sites that overlap with m^6A sites have been identified by comparing PAR-CLIP and single-nucleotide-resolution m^6A sequencing datasets [177]. Computational prediction or experimental measurement of the structures of these sites could serve as a starting point to evaluate whether G3BP1 and G3BP2 act as m^6A 'anti-readers' due to an effect of m^6A on RNA structure.

The mapping of m^6A in human immunodeficiency virus (HIV) RNA has uncovered m^6A sites in the trans-activating response element (TAR) and Rev response element (RRE) RNAs [293–295]. Both TAR and RRE are structured RNAs with bulges, loops, or base triples that perturb A-form RNA helices and influence protein binding [296]. Although the m^6A sites in TAR and RRE have been proposed to enhance binding of the Class I m^6A reader proteins YTHDF1–3 [294, 295], the highly structured nature of these RNAs raises the possibility that m^6A modification impacts RNA structure and influences the binding of structure-dependent Class II m^6A reader proteins. Notably, an m^6A site in the RRE has been proposed to enhance binding of the viral protein Regulator of expression of virion proteins (Rev) and to be critical for nuclear export of viral RNA [293]. It would be interesting to examine the effect of this m^6A site on the structure of the RRE, and to determine whether its impact on Rev binding is mediated by an effect of m^6A on RRE structure. Since the effect of m^6A on Rev binding was only demonstrated *in vivo* [293], it is also possible that m^6A indirectly influences Rev binding, for instance by regulating binding of the human host protein DEAD-box protein 1 (DDX1), which promotes the assembly of the Rev–RRE complex through its interaction with the RRE RNA [297]. Therefore, it might also be worth investigating whether DDX1

binds the RRE in an m^{6}A-dependent manner. Other m^{6}A sites in TAR and RRE have yet to be mapped at single-nucleotide resolution. It would be valuable to map these sites and examine of their effects on the secondary and tertiary structures of the TAR and RRE RNAs. Structural changes induced by m^{6}A could potentially influence the interactions of TAR and RRE with other viral or host proteins.

6.2 Low-Complexity Region Binding to Modified RNAs

In Chapter 3, we discovered and studied a second Class II m^{6}A reader protein [3]. We found that m^{6}A increases the accessibility of its surrounding RNA sequence to bind the protein hnRNPG. Furthermore, hnRNPG binds m^{6}A-containing RNAs through its C-terminal low-complexity region, which self-assembles into large particles *in vitro*. The RGG repeats within the low-complexity region are required for binding to the RNA motif exposed by m^{6}A modification. We further identified 13 191 m^{6}A sites in the transcriptome that regulate RNA–hnRNPG interactions and thereby alter the expression and alternative splicing pattern of target mRNAs. Our results showed that m^{6}A-dependent RNA structural alterations can promote direct binding of m^{6}A-modified RNAs to low-complexity regions.

Although hnRNPG, like hnRNPC, is a Class II m^{6}A reader, it has two distinguishing features: first, hnRNPG binds to a sequence motif that includes the m^{6}A site, and second, hnRNPG uses a low-complexity RGG-containing region to selectively bind m^{6}A-modified RNA. The observation that hnRNPG binds to a sequence motif that includes the m^{6}A site raises the possibility that hnRNPG makes direct contact with the m^{6}A base. If this direct contact occurs, it does not appear to influence the affinity of hnRNPG for the *MALAT1* hairpin that we used in our *in vitro* assays, since mutations that destabilize the *MALAT1* hairpin fully explain the enhanced pull-down of full-length hnRNPG and the increased cross-linking of its low-complexity region. However, the contact with m^{6}A could influence other properties of hnRNPG. In particular, since hnRNPG uses a low-complexity region to bind the *MALAT1* hairpin, one intriguing possibility is that m^{6}A might influence hnRNPG self-

assembly. More generally, direct contact with the m^6A base might alter the conformation of the hnRNPG protein and thereby modulate the interaction of hnRNPG with other proteins or nucleic acids.

A structural approach could be used to assess whether the C-terminal RGG region of hnRNPG makes direct contact with the m^6A base. To reduce protein self-assembly and simplify the system under study, a minimal region of hnRNPG capable of binding the *MALAT1* hairpin should first be identified. To this end, truncated forms of the C-terminal RGG region could be tested for their ability to cross-link to the *MALAT1* RNA. Alternatively, partial digestion of the cross-linked RNA–protein complex, followed by mass spectrometry, could be used to identify which regions of hnRNPG make direct contact with the *MALAT1* RNA. Mutating all three arginines in the RGG repeats of the C-terminal region of hnRNPG (R373F, R377F, R384F) abolished cross-linking to both the nonmethylated and methylated forms of the *MALAT1* hairpin, yet mutating one (R384F) or two (R373F, R377F) of these RGG repeats did not strongly affect cross-linking [3]. While the result with the triple mutant suggests that a minimal hairpin-binding peptide would include the RGG repeats, the results with the single and double mutants suggest that multiple non-overlapping peptides might be capable of binding the *MALAT1* hairpin. Once a minimal hairpin-binding peptide has been identified, structural approaches such as X-ray crystallography or NMR might shed light on the mode through which hnRNPG binds to the nonmethylated and methylated *MALAT1* hairpins, and reveal whether hnRNPG makes direct contact with the m^6A base.

In order to evaluate the effect of RNA–protein and protein–protein interactions on hnRNPG assembly, it would first be necessary to find a reliable approach for measuring the assembly of hnRNPG complexes. To assess the assembly of purified full-length hnRNPG or its C-terminal RGG region *in vitro*, I attempted techniques including negative-staining electron microscopy (EM) and dynamic light scattering (DLS), as described in Chapters 3–4 [3]. However, using these methods, it was difficult to distinguish between true self-assembly of hnRNPG and experimental artifacts. Further optimization and appropriate controls are needed

to more accurately assess the ability of full-length hnRNPG or of its low-complexity region to self-assemble. In particular, the identification of conditions or mutations that alter hnRNPG assembly would provide greater confidence that hnRNPG assembly observed by EM or DLS does not result solely from experimental artifacts. Although the C-terminal low-complexity RGG region of hnRNPG is highly conserved across species (86–100% sequence similarity among tetrapods), this sequence exhibits marked divergence among hnRNPG paralogs, with only 35% sequence conservation in the hnRNPG paralog RNA-binding motif gene on Y chromosome (RBMY) [249]. The divergent low-complexity sequences found in hnRNPG paralogs could be used as a starting point to evaluate the impact of primary sequence on hnRNPG self-assembly. In addition, alternative experimental approaches could be used to investigate hnRNPG assembly. Fluorescent tagging of recombinant hnRNPG would facilitate the visualization of *in vitro* hnRNPG assembly by light microscopy. If hnRNPG self-assembles, this approach could also be used to evaluate the physical properties of hnRNPG complexes. For instance, hnRNPG might form liquid-like droplets, gel-like structures, or fibrous assemblies. Moreover, fluorescent labeling of nonmethylated and methylated RNAs, or of protein binding partners such as the RNAPII CTD, could be used to evaluate whether hnRNPG assemblies can recruit RNA or protein binding partners.

Fluorescent tagging of hnRNPG *in vivo* could help determine whether hnRNPG assembly occurs in cells. Although I observed that hnRNPG forms small nucleoplasmic 'granules' in HEK293T cells (Chapter 4), these 'granules' were observed in fixed cells and were diffraction-limited, so it is unclear whether they were artifacts of the fixation process, and it was difficult to assess their size. Fluorescent tagging of hnRNPG *in vivo* could enable visualization of hnRNPG assembly in live cells. Moreover, super-resolution microscopy could be used to characterize the size, shape, and dynamics of *in vivo* hnRNPG assemblies. Mutations found to affect *in vitro* assembly of hnRNPG could then be introduced in cultured cells to evaluate their effects on *in vivo* hnRNPG assembly and on the cellular functions of hnRNPG, including the regulation of alternative splicing and gene expression. In addition,

Modification type	Modification site	References
Phosphorylation	S88, S329, S332, S352	[299–303]
Dimethylation	R185	[304]
O-glycosylation	S227, S228, S352 (putative sites)	[305]
SUMOylation	K22, K63, K80	[306–308]
Ubiquitylation	K9, K22, K30, K63, K217	[309, 310]

Table 6.1: Selected post-translational modifications in hnRNPG

in vivo fluorescent labeling of RNA or protein binding partners could be used to assess for co-localization with hnRNPG. This approach might also provide insight into the effect of hnRNPG on the assembly of its binding partners. In particular, the transient formation of RNAPII clusters (∼5.1-second lifetime) has been observed by super-resolution microscopy in a human osteosarcoma cell line expressing fluorescently labeled RNAPII catalytic subunit [298]. The depletion of hnRNPG could be used to investigate the effect of hnRNPG on RNAPII assembly in this cell line.

Another question I have not addressed thus far is the effect post-translational modifications on hnRNPG function. Post-translational modifications that have been reported to occur in the hnRNPG protein include serine phosphorylation, arginine dimethylation, *O*-linked glycosylation, SUMOylation (SUMO: small ubiquitin-like modifier), and ubiquitylation (Table 6.1). In particular, phosphorylation has been reported to influence the assembly of certain low-complexity regions [311]. Moreover, arginine methylation, particularly within RGG motifs, can affect RNA–protein and protein–protein interactions, as well as alter protein localization in cells [206]. One difficulty in studying the function of these post-translational modifications is that the enzymes responsible for hnRNPG modification *in vivo* are unknown. Nonetheless, *in vitro* modification of recombinant hnRNPG could be used to evaluate the effect of post-translational modifications on RNA–protein or protein–protein interactions. Likewise, hnRNPG constructs with mutations at modified amino acid residues could provide insights into the roles of these post-translational modifications *in vivo*.

6.3 Co-Transcriptional Functions of RNA Modifications

In Chapter 4, I further studied the cellular functions of the Class II m^6A reader protein hnRNPG. I found that hnRNPG interacts directly with the phosphorylated CTD of transcribing RNAPII by using RGG motifs in its low-complexity region. One of these RGG regions (RGG2) is also used by hnRNPG to recognize an m^6A-modified RNA hairpin [3]. Nonetheless, hnRNPG was able to simultaneously bind to both the CTD and RNA. By introducing mutations in the RRM, RGG1, or RGG2 region of hnRNPG, I found that the RRM and RGG2 regions regulated a larger number of exon splicing events than the RGG1 region. Moreover, the positions of m^6A sites correlated with the direction of splicing regulation for exons regulated by hnRNPG in an RGG2-region-dependent manner. I proposed that hnRNPG co-transcriptionally interacts with both the CTD and nascent RNA to regulate splicing in an m^6A-dependent manner. These results demonstrate how m^6A modifications work in concert with the transcriptional machinery to recruit and regulate the activity of an alternative splicing factor.

Based on the effects of actinomycin D, DRB, and camptothecin on the co-immunoprecipitation of RNAPII with hnRNPG, I proposed that hnRNPG interacts with transcribing RNAPII. However, hnRNPG might further distinguish among transcribing RNAPII enzymes based on whether they are stalled or actively transcribing, or based on which residues of the heptapeptide repeats are phosphorylated. Camptothecin both stalls RNAPII [312] and selectively increases Ser5 phosphorylation (S5P) of the RNAPII CTD [260], and either of these effects might contribute to the observed increase in the interaction between RNAPII and hnRNPG upon camptothecin treatment. The specific enhancement of S5P by camptothecin could be important, since actinomycin D increases phosphorylation of the CTD by CDK9 [313] but decreases co-immunoprecipitation of RNAPII with hnRNPG. While hnRNPG bound to both S2P- and S5P-containing RNAPII in cells and to GST–CTD phosphorylated with either CDK7 or CDK9 *in vitro*, these results do not exclude the possibility that hnRNPG preferentially binds to S5P-CTD, since the CTD contains 52 repeats that can

be differentially phosphorylated [314], and CDK7 and CDK9 have promiscuous phosphorylation activity [315–317]. S5P-containing RNAPII accumulates at the splice sites of actively spliced exons, possibly as a result of RNAPII pausing [39], so a preferential interaction of hnRNPG with stalled, S5P-modified RNAPII could be particularly relevant to the function of hnRNPG in alternative splicing regulation. One approach to determining whether hnRNPG prefers to bind to a specific phosphorylated form of the RNAPII CTD would be to use CTD peptides phosphorylated at specific sites to compete with the phosphorylated GST–CTD for hnRNPG binding *in vitro*. An *in vivo* approach would be to introduce mutations at specific residues in all the heptad repeats of the RNAPII CTD, *e.g.* S2A or S5A, as has been done previously in mammalian cells [24], and then evaluate for the co-immunoprecipitation of RNAPII with hnRNPG. To test the possibility that hnRNPG preferentially interacts with stalled RNAPII, a broader panel of different transcription inhibitors could be used to evaluate for their effect on the RNAPII–hnRNPG interaction [318]. However, transcription inhibitors often affect multiple distinct cellular processes, and it might be difficult to determine which of these effects causes the observed changes in the interaction between RNAPII and hnRNPG. An alternative approach would be to test whether the interaction between hnRNPG and RNAPII depends on transcription rate, by evaluating the interaction of hnRNPG with slow- or fast-transcribing RNAPII mutants *in vivo* [319].

Since pre-mRNA splicing is mostly co-transcriptional, and since hnRNPG likely interacts with the transcribing RNAPII enzyme, I proposed that the regulation of alternative splicing by hnRNPG occurs co-transcriptionally. However, the co-transcriptionality of alternative splicing regulation by hnRNPG has yet to be demonstrated experimentally. I evaluated the effect of hnRNPG on alternative splicing by sequencing polyadenylated RNAs selected from total cellular RNA. To measure co- and post-transcriptional splicing separately, cellular RNA should be separated into distinct fractions, such as chromatin-associated RNA, nucleoplasmic RNA, and cytoplasmic RNA [33, 320]. Changes in co-transcriptional splicing upon depletion or mutation of hnRNPG would be evident in all three fractions (chromatin-associated, nu-

cleoplasmic, and cytoplasmic), whereas changes in post-transcriptional splicing would only be seen in the nucleoplasmic and cytoplasmic fractions. An alternative approach would be to use bromouridine pulse-labeling to measure splicing specifically in nascent RNA [87]. A pulse-labeling approach could also reveal an impact of hnRNPG on splicing rate.

While hnRNPG directly interacts with the phosphorylated CTD of RNAPII, the role of this interaction in the regulation of alternative splicing by hnRNPG, if any, remains to be determined. The role of the RNAPII–hnRNPG interaction in alternative splicing regulation by hnRNPG could be evaluated using RNAPII mutants with a truncated CTD or with point mutations at CTD phosphorylation sites that are important for its interaction with hnRNPG. Although similar experiments have been conducted *in vivo* and *in vitro* [12, 57], a major challenge of such experiments is that perturbations to the CTD would have many effects beyond their effect on hnRNPG binding. Therefore, these experiments could be used to measure splicing in an *in vitro* system or splicing of an extrachromosomal minigene *in vivo*, but they would not be able to measure the effect of the CTD–hnRNPG interaction on the splicing of endogenous genes in their natural chromatin environment. An alternative approach would be to introduce mutations in hnRNPG that abolish its interaction with the CTD without affecting other functions of hnRNPG, including RNA–protein or protein–protein interactions. I attempted this approach by mutating the RGG1 region of hnRNPG in Chapter 4, although it is possible that the RGG1 region has additional unknown functions beyond CTD binding. The small effect of the RGG1 region mutant on alternative splicing might suggest that the interaction of hnRNPG with the CTD is not important for splicing regulation. However, this result is not sufficient to exclude a functional role for the interaction of hnRNPG with the RNAPII CTD. Additional hnRNPG mutants could help gain a better understanding of the requirements for the direct interaction of hnRNPG with the phosphorylated CTD of RNAPII *in vitro*. The introduction of these mutations *in vivo* could then be used to more fully investigate the role of the CTD–hnRNPG interaction in alternative splicing regulation by hnRNPG in cells.

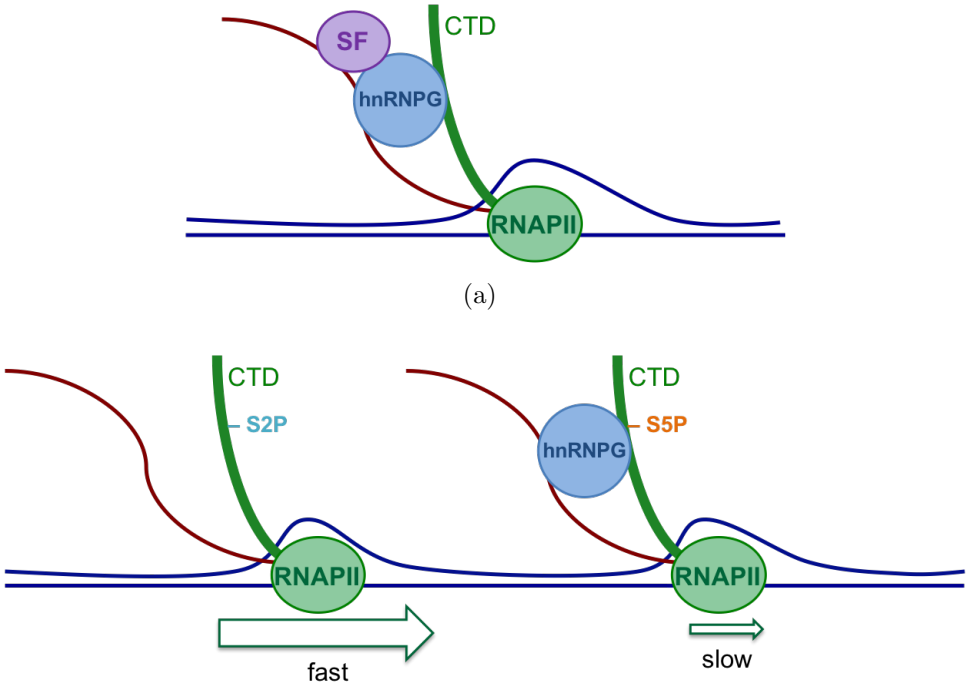


Figure 6.1: Possible mechanisms for the regulation of alternative splicing by hnRNPG. (a) hnRNPG might recruit splicing factors for the m^{6}A-dependent regulation of alternative splicing. (b) hnRNPG might regulate alternative splicing through its effect on the transcription machinery, for instance by affecting RNAPII CTD phosphorylation or elongation rate.

Finally, although I proposed rules governing the direction of splicing regulation by hnRNPG in an m^{6}A-dependent manner, the mechanism by which hnRNPG might regulate alternative splicing has yet to be elucidated. One possible mechanism would be the recruitment of splicing factors by hnRNPG (Figure 6.1(a)), similar to m^{6}A-dependent alternative splicing regulation by YTHDC1 [118]. In fact, hnRNPG is known to interact with the alternative splicing factor transformer 2 β homolog (Tra2β1) [242, 261, 265]. High-throughput studies further suggest that hnRNPG might interact with spliceosome components, including small nuclear ribonucleoprotein U1 subunit 70 (SNRNP70) and splicing factor 3B (SF3B) [321, 322], raising the possibility that hnRNPG recruits or modulates the activity of the spliceosome to regulate alternative splicing. Another possibility is that hnRNPG affects splicing through an effect on the transcription machinery (Figure 6.1(b)). For instance,

hnRNPG might influence the phosphorylation state or elongation rate of RNAPII, as has been demonstrated for the human brahma homolog (Brm), which is a subunit of the mating-type switch/sucrose non-fermenting (SWI/SNF) complex [56]. hnRNPG could affect CTD phosphorylation through the recruitment of kinases or phosphatases, and changes in the phosphorylation state of the RNAPII CTD could in turn influence spliceosome recruitment or transcription elongation rate [11]. Moreover, splicing is kinetically coupled: alternative splicing involves competition between splice sites, and transcription elongation rate can impact alternative splicing by affecting the lag time between the transcription of competing splice sites [31]. A possible effect of hnRNPG on the kinetics of CTD phosphorylation could be evaluated by adding recombinant hnRNPG to *in vitro* CTD phosphorylation reactions and measuring CTD phosphorylation over time. In principle, *in vivo* phosphorylation of the RNAPII CTD could be evaluated by using Western blots to detect different phosphorylated forms of the CTD upon depletion of hnRNPG. However, hnRNPG is unlikely to strongly affect global patterns in the phosphorylation state of the RNAPII CTD, given that hnRNPG only interacts with a small fraction of all RNAPII enzymes in the cell. Therefore, chromatin immunoprecipitation (ChIP) of specific phosphorylated forms of RNAPII might be useful to detect local changes in CTD phosphorylation, for instance within genes that are regulated by hnRNPG. RNAPII ChIP experiments could also reveal an effect of hnRNPG on RNAPII pausing or elongation rate. This approach was used to reveal an effect of the RNA-binding protein Fused in Sarcoma (FUS) on RNAPII CTD phosphorylation [323].

6.4 Transcriptome-Wide Distribution of RNA Modifications

The abundant RNA modification pseudouridine (Ψ) has been mapped transcriptome-wide by chemically modifying pseudouridines with carbodiimide and detecting the resulting reverse transcription stops in high-throughput sequencing. However, these methods have limited sensitivity and specificity, in part due to the use of reverse transcription stops. In Chapter 5, I sought to use mutations rather than only stops in sequencing data to identify pseu-

douridine sites [5]. First, I found reverse transcription conditions that allow read-through of pseudouridine that has been chemically modified with N-cyclohexyl N'-(2-morpholinoethyl) carbodiimide (CMC). Next, we showed that pseudouridines in CMC-treated human ribosomal RNA (rRNA) have context-dependent mutation and stop rates in high-throughput sequencing libraries prepared using these reverse transcription conditions. Furthermore, accounting for the context-dependence of mutation and stop rates can enhance the detection of pseudouridine sites. Similar approaches could contribute to the sequencing-based detection of many RNA modifications.

Most of the pseudouridine sites in human rRNA have previously been mapped using low-throughput methods [160, 285]. By focusing on rRNA, I was able to evaluate the sensitivity and specificity of our context-dependent combined mutation- and stop-based approach for pseudouridine detection. Although the modifications in rRNA have been well-studied, it remains possible that some pseudouridine sites in rRNA have yet to be discovered. In fact, high-throughput sequencing methods have recently revealed previously unknown pseudouridine sites in rRNA, some of which were also detected in our sequencing data [68, 179]. Thus, it is also possible that some of our 'false positive' pseudouridine sites are actually previously unknown pseudouridine sites in rRNA. Moreover, inducible pseudouridine sites have been reported to occur in rRNA [66, 68]. By subjecting cells to different environmental conditions and using our sequencing approach, we might be able to detect even more inducible pseudouridine sites in rRNA.

Despite the potential for the detection of novel pseudouridine sites in rRNA, the goal of this project was ultimately to develop a mutation- and stop-based pseudouridine detection method that could be applied transcriptome-wide. However, the application of our method to mRNA faces major challenges. Although my reverse transcription conditions increased read-through of a CMC-modified pseudouridine (CMC-Ψ) in a synthetic RNA oligo to $\sim$80%, rRNA sequencing yielded a higher stop rate and lower mutation rate than expected based on this result. Therefore, our method remains highly reliant on stop rates to detect pseu-

douridine sites, resulting in many of the same limitations to sensitivity and specificity that apply to other pseudouridine sequencing methods [184]. Since pseudouridine sites in mRNA have much lower modification fractions than pseudouridine sites in rRNA [179], even lower mutation rates would be observed for pseudouridine sites in mRNA. Furthermore, since individual mRNA species are much less abundant than rRNA, it would be impractical to achieve sequence coverage that is high enough to measure very low mutation rates in mRNA.

The discrepancy between the rates of reverse transcription read-through for synthetic and biological substrates could be due to the sensitivity of the reverse transcriptase (RT) enzyme to reaction conditions. Optimization of reverse transcription conditions for cellular RNA could potentially lead to a much higher read-through rate. We also found that the mutation and stop rates at CMC-Ψ depended strongly on the +1 nucleotide context. However, this context-dependence does not explain why reverse transcription of human rRNA yielded a lower read-through rate than reverse transcription of a synthetic RNA, since the synthetic RNA we used had a +1 context (+1 U) that was associated with a low mutation rate and high stop rate. Nonetheless, it remains possible that a broader sequence context, *e.g.* the −2, −1, and +2 nucleotides, might contribute to the higher read-through rate observed for the synthetic RNA. Moreover, a more thorough understanding of context-dependence could further improve the mutation- and stop-based detection of pseudouridine sites. This context-dependence could be further investigated by reverse transcribing standard oligos in which CMC-Ψ is flanked by different combinations of nucleotide bases, and then sequencing the resulting cDNA to determine the mutation rates, stop rates, and mutation signatures observed for each nucleotide context.

Another factor that likely lowered the observed mutation rates is a tendency of HIV RT to incorporate A opposite of CMC-Ψ, due to a preference of HIV RT for incorporating A or G during non-templated synthesis. It might be possible to increase the mutation rate at CMC-Ψ by engineering an HIV RT enzyme with different nucleotide preferences. Alternatively, other RT enzymes could be assayed for their ability not only to read through CMC-Ψ, but

also to introduce mutations at the CMC-Ψ site. The use of a less bulky carbodiimide to modify pseudouridine could also contribute to increased read-through of carbodiimide-modified pseudouridine. In addition, the sensitivity of pseudouridine detection in mRNA could be improved by combining our mutation- and stop-based method with the prior enrichment of carbodiimide-modified RNAs, for instance by using the biotinylated carbodiimide used in CeU-seq [179]. On the other hand, enriching for carbodiimide-modified RNAs would also prevent the quantitative measurement of pseudouridine modification fraction. Ultimately, the goal is to develop a method that not only has high sensitivity, high specificity, and single-nucleotide resolution, but also can quantitatively measure modification fraction. While we are still far from achieving these goals, mutation rate and context-dependence are important considerations to take into account as we work toward developing better methods for the high-throughput detection of pseudouridines and other RNA modifications.

6.5 Conclusion

The work I have presented in this thesis spans diverse topics and experimental approaches in the study of RNA modifications. In my study of the abundant mRNA modification m^6A, I started by investigating how m^6A alters the structure of an RNA hairpin to increase binding of an m^6A reader protein. Next, we identified another m^6A reader protein that binds to m^6A-containing RNA through a low-complexity region, and then I studied the function of this protein in co-transcriptional gene regulation. I also studied another abundant RNA modification, pseudouridine, this time using rRNA as a model system to identify patterns in high-throughput sequencing. Further studies based on this work have the potential to improve the transcriptome-wide mapping of RNA modifications, and to expand our understanding of the impact of RNA modifications on RNA structure, protein binding, and co- and post-transcriptional gene regulation.

References

[1] Katherine I. Zhou, Marc Parisien, Qing Dai, Nian Liu, Luda Diatchenko, Joseph R. Sachleben, and Tao Pan. N^6-Methyladenosine Modification in a Long Noncoding RNA Hairpin Predisposes Its Conformation to Protein Binding. *Journal of Molecular Biology*, 428(5 Pt A):822–833, February 2016.

[2] Katherine I. Zhou and Tao Pan. Structures of the m^6A Methyltransferase Complex: Two Subunits with Distinct but Coordinated Roles. *Molecular Cell*, 63(2):183–185, 2016.

[3] Nian Liu, Katherine I. Zhou, Marc Parisien, Qing Dai, Luda Diatchenko, and Tao Pan. N^6-methyladenosine alters RNA structure to regulate binding of a low-complexity protein. *Nucleic Acids Research*, 45(10):6051–6063, June 2017.

[4] Katherine I. Zhou and Tao Pan. An additional class of m^6A readers. *Nature Cell Biology*, 20(3):230–232, March 2018.

[5] Katherine I. Zhou, Wesley C. Clark, David W. Pan, Matthew J. Eckwahl, Qing Dai, and Tao Pan. Pseudouridines have context-dependent mutation and stop rates in high-throughput sequencing. *RNA biology*, DOI: 10.1080/15476286.2018.1462654, May 2018.

[6] Gal Haimovich, Mordechai Choder, Robert H. Singer, and Tatjana Trcek. The fate of the messenger is pre-determined: a new model for regulation of gene expression. *Biochimica Et Biophysica Acta*, 1829(6-7):643–653, July 2013.

[7] Maija Slaidina and Ruth Lehmann. Translational control in germline stem cell development. *The Journal of Cell Biology*, 207(1):13–21, October 2014.

[8] Tomas S. Andreani, Taichi Q. Itoh, Evrim Yildirim, Dae-Sung Hwangbo, and Ravi Allada. Genetics of Circadian Rhythms. *Sleep Medicine Clinics*, 10(4):413–421, December 2015.

[9] J. P. Herman, J. M. McKlveen, M. B. Solomon, E. Carvalho-Netto, and B. Myers. Neural regulation of the stress response: glucocorticoid feedback mechanisms. *Brazilian Journal of Medical and Biological Research = Revista Brasileira De Pesquisas Medicas E Biologicas*, 45(4):292–298, April 2012.

[10] Kevin M. Harlen and L. Stirling Churchman. The code and beyond: transcription regulation by the RNA polymerase II carboxy-terminal domain. *Nature Reviews. Molecular Cell Biology*, 18(4):263–273, 2017.

[11] Jing-Ping Hsin and James L. Manley. The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes & Development*, 26(19):2119–2137, October 2012.

[12] Manuel J. Muñoz, Manuel de la Mata, and Alberto R. Kornblihtt. The carboxy terminal domain of RNA polymerase II and alternative splicing. *Trends in Biochemical Sciences*, 35(9):497–504, September 2010.

[13] Rob D. Chapman, Martin Heidemann, Corinna Hintermair, and Dirk Eick. Molecular evolution of the RNA polymerase II CTD. *Trends in genetics: TIG*, 24(6):289–296, June 2008.

[14] Dirk Eick and Matthias Geyer. The RNA polymerase II carboxy-terminal domain (CTD) code. *Chemical Reviews*, 113(11):8456–8490, November 2013.

[15] Koon Ho Wong, Yi Jin, and Kevin Struhl. TFIIH phosphorylation of the Pol II CTD stimulates mediator dissociation from the preinitiation complex and promoter escape. *Molecular Cell*, 54(4):601–612, May 2014.

[16] Zhuoyu Ni, Abbie Saunders, Nicholas J. Fuda, Jie Yao, Jose-Ramon Suarez, Watt W. Webb, and John T. Lis. P-TEFb is critical for the maturation of RNA polymerase II into productive elongation in vivo. *Molecular and Cellular Biology*, 28(3):1161–1170, February 2008.

[17] Hua Lu, Osvaldo Flores, Roberto Weinmann, and Danny Reinberg. The nonphosphorylated form of RNA polymerase II preferentially associates with the preinitiation complex. *Proceedings of the National Academy of Sciences of the United States of America*, 88(22):10004–10008, November 1991.

[18] Stephanie C. Schroeder, Beate Schwer, Stewart Shuman, and David Bentley. Dynamic association of capping enzymes with transcribing RNA polymerase II. *Genes & Development*, 14(19):2435–2440, October 2000.

[19] Heidi Sutherland and Wendy A. Bickmore. Transcription factories: gene expression in unions? *Nature Reviews. Genetics*, 10(7):457–466, July 2009.

[20] Hemali P. Phatnani and Arno L. Greenleaf. Phosphorylation and functions of the RNA polymerase II CTD. *Genes & Development*, 20(21):2922–2936, November 2006.

[21] Mark Meininghaus, Rob D. Chapman, Manuela Horndasch, and Dirk Eick. Conditional expression of RNA polymerase II in mammalian cells. Deletion of the carboxyl-terminal domain of the large subunit affects early steps in transcription. *The Journal of Biological Chemistry*, 275(32):24375–24382, August 2000.

[22] Andrew Emili, Michael Shales, Susan McCracken, Weijun Xie, Philip W. Tucker, Ryuji Kobayashi, Benjamin J. Blencowe, and C. James Ingles. Splicing and transcription-associated proteins PSF and p54nrb/nonO bind to the RNA polymerase II CTD. *RNA (New York, N.Y.)*, 8(9):1102–1111, September 2002.

[23] Charles J. David, Alex R. Boyne, Scott R. Millhouse, and James L. Manley. The RNA polymerase II C-terminal domain promotes splicing activation through recruitment of a U2AF65-Prp19 complex. *Genes & Development*, 25(9):972–983, May 2011.

[24] Bo Gu, Dirk Eick, and Olivier Bensaude. CTD serine-2 plays a critical role in splicing and termination factor recruitment to RNA polymerase II in vivo. *Nucleic Acids Research*, 41(3):1591–1603, February 2013.

[25] Seong Hoon Ahn, Minkyu Kim, and Stephen Buratowski. Phosphorylation of serine 2 within the RNA polymerase II C-terminal domain couples transcription and $3'$ end processing. *Molecular Cell*, 13(1):67–76, January 2004.

[26] David Hollingworth, Christian G. Noble, Ian A. Taylor, and Andres Ramos. RNA polymerase II CTD phosphopeptides compete with RNA for the interaction with Pcf11. *RNA (New York, N.Y.)*, 12(4):555–560, April 2006.

[27] Marilyn L. West and Jeffry L. Corden. Construction and analysis of yeast RNA polymerase II CTD deletion and substitution mutations. *Genetics*, 140(4):1223–1233, August 1995.

[28] Beate Schwer and Stewart Shuman. Deciphering the RNA polymerase II CTD code in fission yeast. *Molecular Cell*, 43(2):311–318, July 2011.

[29] Kevin Ryan, Kanneganti G. K. Murthy, Syuzo Kaneko, and James L. Manley. Requirements of the RNA polymerase II C-terminal domain for reconstituting pre-mRNA $3'$ cleavage. *Molecular and Cellular Biology*, 22(6):1684–1692, March 2002.

[30] Bede Portz, Feiyue Lu, Eric B. Gibbs, Joshua E. Mayfield, M. Rachel Mehaffey, Yan Jessie Zhang, Jennifer S. Brodbelt, Scott A. Showalter, and David S. Gilmour. Structural heterogeneity in the intrinsically disordered RNA polymerase II C-terminal domain. *Nature Communications*, 8:15231, May 2017.

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

[32] Fernando Carrillo Oesterreich, Lydia Herzel, Korinna Straube, Katja Hujer, Jonathon Howard, and Karla M. Neugebauer. Splicing of Nascent RNA Coincides with Intron Exit from RNA Polymerase II. *Cell*, 165(2):372–381, April 2016.

[33] Dev M. Bhatt, Amy Pandya-Jones, Ann-Jay Tong, Iros Barozzi, Michelle M. Lissner, Gioacchino Natoli, Douglas L. Black, and Stephen T. Smale. Transcript dynamics of proinflammatory genes revealed by sequence analysis of subcellular RNA fractions. *Cell*, 150(2):279–290, July 2012.

[34] Hagen Tilgner, David G. Knowles, Rory Johnson, Carrie A. Davis, Sudipto Chakrabortty, Sarah Djebali, João Curado, Michael Snyder, Thomas R. Gingeras, and Roderic Guigó. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Research*, 22(9):1616–1625, September 2012.

[35] Yevgenia L. Khodor, Jerome S. Menet, Michael Tolan, and Michael Rosbash. Cotranscriptional splicing efficiency differs dramatically between Drosophila and mouse. *RNA (New York, N.Y.)*, 18(12):2174–2186, December 2012.

[36] Yevgenia L. Khodor, Joseph Rodriguez, Katharine C. Abruzzi, Chih-Hang Anthony Tang, Michael T. Marr, and Michael Rosbash. Nascent-seq indicates widespread cotranscriptional pre-mRNA splicing in Drosophila. *Genes & Development*, 25(23):2502–2512, December 2011.

[37] L. Stirling Churchman and Jonathan S. Weissman. Native elongating transcript sequencing (NET-seq). *Current Protocols in Molecular Biology*, Chapter 4:Unit 4.14.1–17, April 2012.

[38] Takayuki Nojima, Tomás Gomes, Ana Rita Fialho Grosso, Hiroshi Kimura, Michael J. Dye, Somdutta Dhir, Maria Carmo-Fonseca, and Nicholas J. Proudfoot. Mammalian NET-Seq Reveals Genome-wide Nascent Transcription Coupled to RNA Processing. *Cell*, 161(3):526–540, April 2015.

[39] Andreas Mayer, Julia di Iulio, Seth Maleri, Umut Eser, Jeff Vierstra, Alex Reynolds, Richard Sandstrom, John A. Stamatoyannopoulos, and L. Stirling Churchman. Native elongating transcript sequencing reveals human transcriptional activity at nucleotide resolution. *Cell*, 161(3):541–554, April 2015.

[40] Kevin M. Harlen, Kristine L. Trotta, Erin E. Smith, Mohammad M. Mosaheb, Stephen M. Fuchs, and L. Stirling Churchman. Comprehensive RNA Polymerase II Interactomes Reveal Distinct and Varied Roles for Each Phospho-CTD Residue. *Cell Reports*, 15(10):2147–2158, 2016.

[41] Qun Pan, Ofer Shai, Leo J. Lee, Brendan J. Frey, and Benjamin J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40(12):1413–1415, December 2008.

[42] Arianne J. Matlin, Francis Clark, and Christopher W. J. Smith. Understanding alternative splicing: towards a cellular code. *Nature Reviews Molecular Cell Biology*, 6(5):386–398, May 2005.

[43] Xiang-Dong Fu and Manuel Ares. Context-dependent control of alternative splicing by RNA-binding proteins. *Nature Reviews Genetics*, 15(10):689–701, August 2014.

[44] Jennifer C. Long and Javier F. Caceres. The SR protein family of splicing factors: master regulators of gene expression. *The Biochemical Journal*, 417(1):15–27, January 2009.

[45] Shalini Sharma, Lori A. Kohlstaedt, Andrey Damianov, Donald C. Rio, and Douglas L. Black. Polypyrimidine tract binding protein controls the transition from exon definition to an intron defined spliceosome. *Nature Structural & Molecular Biology*, 15(2):183–191, February 2008.

[46] Shalini Sharma, Christophe Maris, Frédéric H.-T. Allain, and Douglas L. Black. U1 snRNA directly interacts with polypyrimidine tract-binding protein during splicing repression. *Molecular Cell*, 41(5):579–588, March 2011.

[47] Ni-Ting Chiou, Ganesh Shankarling, and Kristen W. Lynch. hnRNP L and hnRNP A1 induce extended U1 snRNA interactions with an exon to repress spliceosome assembly. *Molecular Cell*, 49(5):972–982, March 2013.

[48] Florian C. Oberstrass, Sigrid D. Auweter, Michèle Erat, Yann Hargous, Anke Henning, Philipp Wenter, Luc Reymond, Batoul Amir-Ahmady, Stefan Pitsch, Douglas L. Black, and Frédéric H.-T. Allain. Structure of PTB bound to RNA: specific binding and implications for splicing regulation. *Science (New York, N.Y.)*, 309(5743):2054–2057, September 2005.

[49] Rebeca Martinez-Contreras, Jean-François Fisette, Faiz-ul Hassan Nasim, Richard Madden, Mélanie Cordeau, and Benoit Chabot. Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing. *PLoS biology*, 4(2):e21, February 2006.

[50] Arneet L. Saltzman, Qun Pan, and Benjamin J. Blencowe. Regulation of alternative splicing by the core spliceosomal machinery. *Genes & Development*, 25(4):373–384, February 2011.

[51] Shatakshi Pandit, Yu Zhou, Lily Shiue, Gabriela Coutinho-Mansfield, Hairi Li, Jinsong Qiu, Jie Huang, Gene W. Yeo, Manuel Ares, and Xiang-Dong Fu. Genome-wide analysis reveals SR protein cooperation and competition in regulated splicing. *Molecular Cell*, 50(2):223–235, April 2013.

[52] Miriam Llorian, Schraga Schwartz, Tyson A. Clark, Dror Hollander, Lit-Yeen Tan, Rachel Spellman, Adele Gordon, Anthony C. Schweitzer, Pierre de la Grange, Gil Ast, and Christopher W. J. Smith. Position-dependent alternative splicing activity revealed by global profiling of alternative splicing events regulated by PTB. *Nature Structural & Molecular Biology*, 17(9):1114–1123, September 2010.

[53] Stephanie C. Huelga, Anthony Q. Vu, Justin D. Arnold, Tiffany Y. Liang, Patrick P. Liu, Bernice Y. Yan, John Paul Donohue, Lily Shiue, Shawn Hoon, Sydney Brenner, Manuel Ares, and Gene W. Yeo. Integrative genome-wide analysis reveals cooperative regulation of alternative splicing by hnRNP proteins. *Cell reports*, 1(2):167–178, February 2012.

[54] Alberto R. Kornblihtt, Ignacio E. Schor, Mariano Alló, Gwendal Dujardin, Ezequiel Petrillo, and Manuel J. Muñoz. Alternative splicing: a pivotal step between eukaryotic transcription and translation. *Nature Reviews. Molecular Cell Biology*, 14(3):153–165, March 2013.

[55] Gwendal Dujardin, Celina Lafaille, Manuel de la Mata, Luciano E. Marasco, Manuel J. Muñoz, Catherine Le Jossic-Corcos, Laurent Corcos, and Alberto R. Kornblihtt. How slow RNA polymerase II elongation favors alternative exon skipping. *Molecular Cell*, 54(4):683–690, May 2014.

[56] Eric Batsché, Moshe Yaniv, and Christian Muchardt. The human SWI/SNF subunit Brm is a regulator of alternative splicing. *Nature Structural & Molecular Biology*, 13(1):22–29, January 2006.

[57] Manuel de la Mata and Alberto R. Kornblihtt. RNA polymerase II C-terminal domain mediates regulation of alternative splicing by SRp20. *Nature Structural & Molecular Biology*, 13(11):973–980, November 2006.

[58] Ian A. Roundtree, Molly E. Evans, Tao Pan, and Chuan He. Dynamic RNA Modifications in Gene Expression Regulation. *Cell*, 169(7):1187–1200, June 2017.

[59] Ronald Desrosiers, Karen Friderici, and Fritz Rottman. Identification of methylated nucleosides in messenger RNA from Novikoff hepatoma cells. *Proceedings of the National Academy of Sciences*, 71(10):3971–3975, 1974.

[60] Dan Dominissini, Sharon Moshitch-Moshkovitz, Schraga Schwartz, Mali Salmon-Divon, Lior Ungar, Sivan Osenberg, Karen Cesarkas, Jasmine Jacob-Hirsch, Ninette Amariglio, Martin Kupiec, Rotem Sorek, and Gideon Rechavi. Topology of the human and mouse m^6A RNA methylomes revealed by m^6A-seq. *Nature*, 485(7397):201–206, April 2012.

[61] Kate D. Meyer, Yogesh Saletore, Paul Zumbo, Olivier Elemento, Christopher E. Mason, and Samie R. Jaffrey. Comprehensive Analysis of mRNA Methylation Reveals Enrichment in $3'$ UTRs and near Stop Codons. *Cell*, 149(7):1635–1646, June 2012.

[62] Dan Dominissini, Sigrid Nachtergaele, Sharon Moshitch-Moshkovitz, Eyal Peer, Nitzan Kol, Moshe Shay Ben-Haim, Qing Dai, Ayelet Di Segni, Mali Salmon-Divon, Wesley C. Clark, Guanqun Zheng, Tao Pan, Oz Solomon, Eran Eyal, Vera Hershkovitz, Dali Han, Louis C. Doré, Ninette Amariglio, Gideon Rechavi, and Chuan He. The dynamic N1-methyladenosine methylome in eukaryotic messenger RNA. *Nature*, 530(7591):441–446, February 2016.

[63] Jeffrey E. Squires, Hardip R. Patel, Marco Nousch, Tennille Sibbritt, David T. Humphreys, Brian J. Parker, Catherine M. Suter, and Thomas Preiss. Widespread occurrence of 5-methylcytosine in human coding and non-coding RNA. *Nucleic Acids Research*, 40(11):5023–5033, June 2012.

[64] Schraga Schwartz, Sudeep D. Agarwala, Maxwell R. Mumbach, Marko Jovanovic, Philipp Mertins, Alexander Shishkin, Yuval Tabach, Tarjei S. Mikkelsen, Rahul Satija, Gary Ruvkun, Steven A. Carr, Eric S. Lander, Gerald R. Fink, and Aviv Regev. High-Resolution Mapping Reveals a Conserved, Widespread, Dynamic mRNA Methylation Program in Yeast Meiosis. *Cell*, 155(6):1409–1421, December 2013.

[65] Benjamin Delatte, Fei Wang, Long Vo Ngoc, Evelyne Collignon, Elise Bonvin, Rachel Deplus, Emilie Calonne, Bouchra Hassabi, Pascale Putmans, Stephan Awe, Collin Wetzel, Judith Kreher, Romuald Soin, Catherine Creppe, Patrick A. Limbach, Cyril Gueydan, Véronique Kruys, Alexander Brehm, Svetlana Minakhina, Matthieu Defrance, Ruth Steward, and François Fuks. RNA biochemistry. Transcriptome-wide

distribution and function of RNA hydroxymethylcytosine. *Science (New York, N.Y.)*, 351(6270):282–285, January 2016.

[66] Schraga Schwartz, Douglas A. Bernstein, Maxwell R. Mumbach, Marko Jovanovic, Rebecca H. Herbst, Brian X. León-Ricardo, Jesse M. Engreitz, Mitchell Guttman, Rahul Satija, Eric S. Lander, Gerald Fink, and Aviv Regev. Transcriptome-wide Mapping Reveals Widespread Dynamic-Regulated Pseudouridylation of ncRNA and mRNA. *Cell*, 159(1):148–162, September 2014.

[67] Alexander F. Lovejoy, Daniel P. Riordan, and Patrick O. Brown. Transcriptome-wide mapping of pseudouridines: pseudouridine synthases modify specific mRNAs in S. cerevisiae. *PloS One*, 9(10):e110799, 2014.

[68] Thomas M. Carlile, Maria F. Rojas-Duran, Boris Zinshteyn, Hakyung Shin, Kristen M. Bartoli, and Wendy V. Gilbert. Pseudouridine profiling reveals regulated mRNA pseudouridylation in yeast and human cells. *Nature*, 515(7525):143–146, September 2014.

[69] Guifang Jia, Ye Fu, Xu Zhao, Qing Dai, Guanqun Zheng, Ying Yang, Chengqi Yi, Tomas Lindahl, Tao Pan, Yun-Gui Yang, and Chuan He. N6-Methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nature Chemical Biology*, 7(12):885–887, October 2011.

[70] Tao Pan. N^6-methyl-adenosine modification in messenger and long non-coding RNA. *Trends in Biochemical Sciences*, 38(4):204–209, April 2013.

[71] Bastian Linder, Anya V. Grozhik, Anthony O. Olarerin-George, Cem Meydan, Christopher E. Mason, and Samie R. Jaffrey. Single-nucleotide-resolution mapping of m^6A and m^6A_m throughout the transcriptome. *Nature Methods*, 12(8):767–772, June 2015.

[72] Nian Liu, Marc Parisien, Qing Dai, Guanqun Zheng, Chuan He, and Tao Pan. Probing N^6-methyladenosine RNA modification status at single nucleotide resolution in mRNA and long noncoding RNA. *RNA*, 19(12):1848–1856, December 2013.

[73] Boxuan Simen Zhao, Ian A. Roundtree, and Chuan He. Post-transcriptional gene regulation by mRNA modifications. *Nature Reviews. Molecular Cell Biology*, 18(1):31–42, 2017.

[74] Jianzhao Liu, Yanan Yue, Dali Han, Xiao Wang, Ye Fu, Liang Zhang, Guifang Jia, Miao Yu, Zhike Lu, Xin Deng, Qing Dai, Weizhong Chen, and Chuan He. A METTL3–METTL14 complex mediates mammalian nuclear RNA N^6-adenosine methylation. *Nature Chemical Biology*, 10(2):93–95, February 2014.

[75] Xiang Wang, Jing Feng, Yuan Xue, Zeyuan Guan, Delin Zhang, Zhu Liu, Zhou Gong, Qiang Wang, Jinbo Huang, Chun Tang, Tingting Zou, and Ping Yin. Structural basis of N^6-adenosine methylation by the METTL3–METTL14 complex. *Nature*, advance online publication, May 2016.

[76] Ping Wang, Katelyn A. Doxtader, and Yunsam Nam. Structural basis for cooperative function of Mettl3 and Mettl14 methyltransferases. *Molecular Cell*, 62, 2016.

[77] Paweł Śledź and Martin Jinek. Structural insights into the molecular mechanism of the m^{6}A writer complex. *eLife*, 5, 2016.

[78] Yang Wang, Yue Li, Julia I. Toth, Matthew D. Petroski, Zhaolei Zhang, and Jing Crystal Zhao. N^6-methyladenosine modification destabilizes developmental regulators in embryonic stem cells. *Nature Cell Biology*, 16(2):191–198, January 2014.

[79] Jinbo Huang, Xu Dong, Zhou Gong, Ling-Yun Qin, Shuai Yang, Yue-Ling Zhu, Xiang Wang, Delin Zhang, Tingting Zou, Ping Yin, and Chun Tang. Solution structure of the RNA recognition domain of METTL3–METTL14 N^6-methyladenosine methyltransferase. *Protein & Cell*, March 2018.

[80] Xiao-Li Ping, Bao-Fa Sun, Lu Wang, Wen Xiao, Xin Yang, Wen-Jia Wang, Samir Adhikari, Yue Shi, Ying Lv, Yu-Sheng Chen, Xu Zhao, Ang Li, Ying Yang, Ujwal Dahal, Xiao-Min Lou, Xi Liu, Jun Huang, Wei-Ping Yuan, Xiao-Fan Zhu, Tao Cheng, Yong-Liang Zhao, Xinquan Wang, Jannie M. Rendtlew Danielsen, Feng Liu, and Yun-Gui Yang. Mammalian WTAP is a regulatory subunit of the RNA N^6-methyladenosine methyltransferase. *Cell Research*, 24(2):177–189, February 2014.

[81] Schraga Schwartz, Maxwell R. Mumbach, Marko Jovanovic, Tim Wang, Karolina Maciag, G. Guy Bushkin, Philipp Mertins, Dmitry Ter-Ovanesyan, Naomi Habib, Davide Cacchiarelli, Neville E. Sanjana, Elizaveta Freinkman, Michael E. Pacold, Rahul Satija, Tarjei S. Mikkelsen, Nir Hacohen, Feng Zhang, Steven A. Carr, Eric S. Lander, and Aviv Regev. Perturbation of m^{6}A Writers Reveals Two Distinct Classes of mRNA Methylation at Internal and 5′ Sites. *Cell Reports*, 8(1):284–296, July 2014.

[82] Yanan Yue, Jun Liu, Xiaolong Cui, Jie Cao, Guanzheng Luo, Zezhou Zhang, Tao Cheng, Minsong Gao, Xiao Shu, Honghui Ma, Fengqin Wang, Xinxia Wang, Bin Shen, Yizhen Wang, Xinhua Feng, Chuan He, and Jianzhao Liu. VIRMA mediates preferential m^{6}A mRNA methylation in 3′ UTR and near stop codon and associates with alternative polyadenylation. *Cell Discovery*, 4:10, 2018.

[83] Deepak P. Patil, Chun-Kan Chen, Brian F. Pickering, Amy Chow, Constanza Jackson, Mitchell Guttman, and Samie R. Jaffrey. m^{6}A RNA methylation promotes XIST-mediated transcriptional repression. *Nature*, 537(7620):369–373, 2016.

[84] Tong Chen, Ya-Juan Hao, Ying Zhang, Miao-Miao Li, Meng Wang, Weifang Han, Yongsheng Wu, Ying Lv, Jie Hao, Libin Wang, Ang Li, Ying Yang, Kang-Xuan Jin, Xu Zhao, Yuhuan Li, Xiao-Li Ping, Wei-Yi Lai, Li-Gang Wu, Guibin Jiang, Hai-Lin Wang, Lisi Sang, Xiu-Jie Wang, Yun-Gui Yang, and Qi Zhou. m^{6}A RNA Methylation Is Regulated by MicroRNAs and Promotes Reprogramming to Pluripotency. *Cell Stem Cell*, 16(3):289–301, March 2015.

[85] Isaia Barbieri, Konstantinos Tzelepis, Luca Pandolfini, Junwei Shi, Gonzalo Millán-Zambrano, Samuel C. Robson, Demetrios Aspris, Valentina Migliori, Andrew J. Bannister, Namshik Han, Etienne De Braekeleer, Hannes Ponstingl, Alan Hendrick, Christopher R. Vakoc, George S. Vassiliou, and Tony Kouzarides. Promoter-bound

METTL3 maintains myeloid leukaemia by m^{6}A-dependent translation control. *Nature*, 552(7683):126–131, December 2017.

[86] Philip Knuckles, Sarah H. Carl, Michael Musheev, Christof Niehrs, Alice Wenger, and Marc Bühler. RNA fate determination through cotranscriptional adenosine methylation and microprocessor binding. *Nature Structural & Molecular Biology*, 24(7):561–569, July 2017.

[87] Annita Louloupi, Evgenia Ntini, Thomas Conrad, and Ulf Andersson Orom. Transient N^6-methyladenosine Transcriptome sequencing reveals a regulatory role of m^{6}A in splicing efficiency. *bioRxiv*, page 242966, January 2018.

[88] Shengdong Ke, Amy Pandya-Jones, Yuhki Saito, John J. Fak, Cathrine Broberg Vagbø, Shay Geula, Jacob H. Hanna, Douglas L. Black, James E. Darnell, and Robert B. Darnell. m^{6}A mRNA modifications are deposited in nascent pre-mRNA and are not required for splicing but do specify cytoplasmic turnover. *Genes & Development*, 31(10):990–1006, 2017.

[89] Boris Slobodin, Ruiqi Han, Vittorio Calderone, Joachim A. F. Oude Vrielink, Fabricio Loayza-Puch, Ran Elkon, and Reuven Agami. Transcription Impacts the Efficiency of mRNA Translation via Co-transcriptional N^6-adenosine Methylation. *Cell*, 169(2):326–337.e12, April 2017.

[90] Kathryn E. Pendleton, Beibei Chen, Kuanqing Liu, Olga V. Hunter, Yang Xie, Benjamin P. Tu, and Nicholas K. Conrad. The U6 snRNA m^{6}A Methyltransferase METTL16 Regulates SAM Synthetase Intron Retention. *Cell*, 169(5):824–835.e14, May 2017.

[91] Shay Geula, Sharon Moshitch-Moshkovitz, Dan Dominissini, Abed AlFatah Mansour, Nitzan Kol, Mali Salmon-Divon, Vera Hershkovitz, Eyal Peer, Nofar Mor, Yair S. Manor, Moshe Shay Ben-Haim, Eran Eyal, Sharon Yunger, Yishay Pinto, Diego Adhemar Jaitin, Sergey Viukov, Yoach Rais, Vladislav Krupalnik, Elad Chomsky, Mirie Zerbib, Itay Maza, Yoav Rechavi, Rada Massarwa, Suhair Hanna, Ido Amit, Erez Y. Levanon, Ninette Amariglio, Noam Stern-Ginossar, Noa Novershtern, Gideon Rechavi, and Jacob H. Hanna. m^{6}A mRNA methylation facilitates resolution of naïve pluripotency toward differentiation. *Science (New York, N.Y.)*, 347(6225):1002–1006, February 2015.

[92] Tina Lence, Junaid Akhtar, Marc Bayer, Katharina Schmid, Laura Spindler, Cheuk Hei Ho, Nastasja Kreim, Miguel A. Andrade-Navarro, Burkhard Poeck, Mark Helm, and Jean-Yves Roignant. m^{6}A modulates neuronal functions and sex determination in Drosophila. *Nature*, 540(7632):242–247, 2016.

[93] Irmgard U. Haussmann, Zsuzsanna Bodi, Eugenio Sanchez-Moran, Nigel P. Mongan, Nathan Archer, Rupert G. Fray, and Matthias Soller. m^{6}A potentiates Sxl alternative pre-mRNA splicing for robust Drosophila sex determination. *Nature*, 540(7632):301–304, 2016.

[94] Pedro J. Batista, Benoit Molinie, Jinkai Wang, Kun Qu, Jiajing Zhang, Lingjie Li, Donna M. Bouley, Ernesto Lujan, Bahareh Haddad, Kaveh Daneshvar, Ava C. Carter, Ryan A. Flynn, Chan Zhou, Kok-Seong Lim, Peter Dedon, Marius Wernig, Alan C. Mullen, Yi Xing, Cosmas C. Giallourakis, and Howard Y. Chang. m^{6}A RNA Modification Controls Cell Fate Transition in Mammalian Embryonic Stem Cells. *Cell Stem Cell*, 15(6):707–719, December 2014.

[95] Yang Wang, Yue Li, Minghui Yue, Jun Wang, Sandeep Kumar, Robert J. Wechsler-Reya, Zhaolei Zhang, Yuya Ogawa, Manolis Kellis, Gregg Duester, and Jing Crystal Zhao. N^6-methyladenosine RNA modification regulates embryonic neural stem cell self-renewal through histone modifications. *Nature Neuroscience*, 21(2):195–206, February 2018.

[96] Ki-Jun Yoon, Francisca Rojas Ringeling, Caroline Vissers, Fadi Jacob, Michael Pokrass, Dennisse Jimenez-Cyrus, Yijing Su, Nam-Shik Kim, Yunhua Zhu, Lily Zheng, Sunghan Kim, Xinyuan Wang, Louis C. Doré, Peng Jin, Sergi Regot, Xiaoxi Zhuang, Stefan Canzar, Chuan He, Guo-Li Ming, and Hongjun Song. Temporal Control of Mammalian Cortical Neurogenesis by m^{6}A Methylation. *Cell*, 171(4):877–889.e17, November 2017.

[97] Chunxia Zhang, Yusheng Chen, Baofa Sun, Lu Wang, Ying Yang, Dongyuan Ma, Junhua Lv, Jian Heng, Yanyan Ding, Yuanyuan Xue, Xinyan Lu, Wen Xiao, Yun-Gui Yang, and Feng Liu. m^{6}A modulates haematopoietic stem and progenitor cell specification. *Nature*, 549(7671):273–276, 2017.

[98] Jean-Michel Fustin, Masao Doi, Yoshiaki Yamaguchi, Hayashi Hida, Shinichi Nishimura, Minoru Yoshida, Takayuki Isagawa, Masaki Suimye Morioka, Hideaki Kakeya, Ichiro Manabe, and Hitoshi Okamura. RNA-Methylation-Dependent RNA Processing Controls the Speed of the Circadian Clock. *Cell*, 155(4):793–806, November 2013.

[99] Pedro J. Batista. The RNA Modification N^6-methyladenosine and Its Implications in Human Disease. *Genomics, Proteomics & Bioinformatics*, 15(3):154–163, 2017.

[100] Shuibin Lin, Junho Choe, Peng Du, Robinson Triboulet, and Richard I. Gregory. The m^{6}A Methyltransferase METTL3 Promotes Translation in Human Cancer Cells. *Molecular Cell*, 62(3):335–345, May 2016.

[101] Ly P. Vu, Brian F. Pickering, Yuanming Cheng, Sara Zaccara, Diu Nguyen, Gerard Minuesa, Timothy Chou, Arthur Chow, Yogesh Saletore, Matthew MacKay, Jessica Schulman, Christopher Famulare, Minal Patel, Virginia M. Klimek, Francine E. Garrett-Bakelman, Ari Melnick, Martin Carroll, Christopher E. Mason, Samie R. Jaffrey, and Michael G. Kharas. The N^6-methyladenosine (m^{6}a)-forming enzyme METTL3 controls myeloid differentiation of normal hematopoietic and leukemia cells. *Nature Medicine*, 23(11):1369–1376, November 2017.

[102] Hengyou Weng, Huilin Huang, Huizhe Wu, Xi Qin, Boxuan Simen Zhao, Lei Dong, Hailing Shi, Jennifer Skibbe, Chao Shen, Chao Hu, Yue Sheng, Yungui Wang, Mark

Wunderlich, Bin Zhang, Louis C. Dore, Rui Su, Xiaolan Deng, Kyle Ferchen, Chenying Li, Miao Sun, Zhike Lu, Xi Jiang, Guido Marcucci, James C. Mulloy, Jianhua Yang, Zhijian Qian, Minjie Wei, Chuan He, and Jianjun Chen. METTL14 Inhibits Hematopoietic Stem/Progenitor Differentiation and Promotes Leukemogenesis via mRNA m^6A Modification. *Cell Stem Cell*, 22(2):191–205.e9, February 2018.

[103] Guanqun Zheng, John Arne Dahl, Yamei Niu, Peter Fedorcsak, Chun-Min Huang, Charles J. Li, Cathrine B. Vagbø, Yue Shi, Wen-Ling Wang, Shu-Hui Song, Zhike Lu, Ralph P. G. Bosmans, Qing Dai, Ya-Juan Hao, Xin Yang, Wen-Ming Zhao, Wei-Min Tong, Xiu-Jie Wang, Florian Bogdan, Kari Furu, Ye Fu, Guifang Jia, Xu Zhao, Jun Liu, Hans E. Krokan, Arne Klungland, Yun-Gui Yang, and Chuan He. ALKBH5 Is a Mammalian RNA Demethylase that Impacts RNA Metabolism and Mouse Fertility. *Molecular Cell*, 49(1):18–29, October 2013.

[104] Ye Fu, Guifang Jia, Xueqin Pang, Richard N. Wang, Xiao Wang, Charles J. Li, Scott Smemo, Qing Dai, Kathleen A. Bailey, Marcelo A. Nobrega, Ke-Li Han, Qiang Cui, and Chuan He. FTO-mediated formation of N^6-hydroxymethyladenosine and N^6-formyladenosine in mammalian RNA. *Nature Communications*, 4:1798, 2013.

[105] Shui Zou, Joel D. W. Toh, Kendra H. Q. Wong, Yong-Gui Gao, Wanjin Hong, and Esther C. Y. Woon. N^6-Methyladenosine: a conformational marker that regulates the substrate specificity of human demethylases FTO and ALKBH5. *Scientific Reports*, 6:25677, 2016.

[106] Jan Mauer, Xiaobing Luo, Alexandre Blanjoie, Xinfu Jiao, Anya V. Grozhik, Deepak P. Patil, Bastian Linder, Brian F. Pickering, Jean-Jacques Vasseur, Qiuying Chen, Steven S. Gross, Olivier Elemento, Françoise Debart, Megerditch Kiledjian, and Samie R. Jaffrey. Reversible methylation of m^6A_m in the $5'$ cap controls mRNA stability. *Nature*, 541(7637):371–375, 2017.

[107] Christian Dina, David Meyre, Sophie Gallina, Emmanuelle Durand, Antje Körner, Peter Jacobson, Lena M. S. Carlsson, Wieland Kiess, Vincent Vatin, Cecile Lecoeur, Jérome Delplanque, Emmanuel Vaillant, François Pattou, Juan Ruiz, Jacques Weill, Claire Levy-Marchal, Fritz Horber, Natascha Potoczna, Serge Hercberg, Catherine Le Stunff, Pierre Bougnères, Peter Kovacs, Michel Marre, Beverley Balkau, Stéphane Cauchi, Jean-Claude Chèvre, and Philippe Froguel. Variation in FTO contributes to childhood obesity and severe adult obesity. *Nature Genetics*, 39(6):724–726, June 2007.

[108] Timothy M. Frayling, Nicholas J. Timpson, Michael N. Weedon, Eleftheria Zeggini, Rachel M. Freathy, Cecilia M. Lindgren, John R. B. Perry, Katherine S. Elliott, Hana Lango, Nigel W. Rayner, Beverley Shields, Lorna W. Harries, Jeffrey C. Barrett, Sian Ellard, Christopher J. Groves, Bridget Knight, Ann-Marie Patch, Andrew R. Ness, Shah Ebrahim, Debbie A. Lawlor, Susan M. Ring, Yoav Ben-Shlomo, Marjo-Riitta Jarvelin, Ulla Sovio, Amanda J. Bennett, David Melzer, Luigi Ferrucci, Ruth J. F. Loos, Inês Barroso, Nicholas J. Wareham, Fredrik Karpe, Katharine R. Owen, Lon R. Cardon, Mark Walker, Graham A. Hitman, Colin N. A. Palmer, Alex S. F. Doney, Andrew D. Morris, George Davey Smith, Andrew T. Hattersley, and Mark I. McCarthy.

A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science (New York, N.Y.)*, 316(5826):889–894, May 2007.

[109] Angelo Scuteri, Serena Sanna, Wei-Min Chen, Manuela Uda, Giuseppe Albai, James Strait, Samer Najjar, Ramaiah Nagaraja, Marco Orrú, Gianluca Usala, Mariano Dei, Sandra Lai, Andrea Maschio, Fabio Busonero, Antonella Mulas, Georg B. Ehret, Ashley A. Fink, Alan B. Weder, Richard S. Cooper, Pilar Galan, Aravinda Chakravarti, David Schlessinger, Antonio Cao, Edward Lakatta, and Gonçalo R. Abecasis. Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. *PLoS genetics*, 3(7):e115, July 2007.

[110] Scott Smemo, Juan J. Tena, Kyoung-Han Kim, Eric R. Gamazon, Noboru J. Sakabe, Carlos Gómez-Marín, Ivy Aneas, Flavia L. Credidio, Débora R. Sobreira, Nora F. Wasserman, Ju Hee Lee, Vijitha Puviindran, Davis Tam, Michael Shen, Joe Eun Son, Niki Alizadeh Vakili, Hoon-Ki Sung, Silvia Naranjo, Rafael D. Acemel, Miguel Manzanares, Andras Nagy, Nancy J. Cox, Chi-Chung Hui, Jose Luis Gomez-Skarmeta, and Marcelo A. Nóbrega. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*, 507(7492):371–375, March 2014.

[111] Chris Church, Sheena Lee, Eleanor A. L. Bagg, James S. McTaggart, Robert Deacon, Thomas Gerken, Angela Lee, Lee Moir, Jasmin Mecinovi, Mohamed M. Quwailid, Christopher J. Schofield, Frances M. Ashcroft, and Roger D. Cox. A mouse model for the metabolic effects of the human fat mass and obesity associated FTO gene. *PLoS genetics*, 5(8):e1000599, August 2009.

[112] Chris Church, Lee Moir, Fiona McMurray, Christophe Girard, Gareth T. Banks, Lydia Teboul, Sara Wells, Jens C. Brüning, Patrick M. Nolan, Frances M. Ashcroft, and Roger D. Cox. Overexpression of Fto leads to increased food intake and results in obesity. *Nature Genetics*, 42(12):1086–1092, December 2010.

[113] Xu Zhao, Ying Yang, Bao-Fa Sun, Yue Shi, Xin Yang, Wen Xiao, Ya-Juan Hao, Xiao-Li Ping, Yu-Sheng Chen, Wen-Jia Wang, Kang-Xuan Jin, Xing Wang, Chun-Min Huang, Yu Fu, Xiao-Meng Ge, Shu-Hui Song, Hyun Seok Jeong, Hiroyuki Yanagisawa, Yamei Niu, Gui-Fang Jia, Wei Wu, Wei-Min Tong, Akimitsu Okamoto, Chuan He, Jannie M. Rendtlew Danielsen, Xiu-Jie Wang, and Yun-Gui Yang. FTO-dependent demethylation of N^6-methyladenosine regulates mRNA splicing and is required for adipogenesis. *Cell Research*, 24(12):1403–1419, December 2014.

[114] Jocelyn Widagdo, Qiong-Yi Zhao, Marie-Jeanne Kempen, Men Chee Tan, Vikram S. Ratnu, Wei Wei, Laura Leighton, Paola A. Spadaro, Janette Edson, Victor Anggono, and Timothy W. Bredy. Experience-Dependent Accumulation of N^6-Methyladenosine in the Prefrontal Cortex Is Associated with Memory Processes in Mice. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 36(25):6771–6777, 2016.

[115] Marek Bartosovic, Helena Covelo Molares, Pavlina Gregorova, Dominika Hrossova, Grzegorz Kudla, and Stepanka Vanacova. N^6-methyladenosine demethylase FTO targets pre-mRNAs and regulates alternative splicing and 3′-end processing. *Nucleic Acids Research*, 45(19):11356–11370, November 2017.

[116] Jun Yu, Mengxian Chen, Haijiao Huang, Junda Zhu, Huixue Song, Jian Zhu, Jaewon Park, and Sheng-Jian Ji. Dynamic m^6A modification regulates local translation of mRNA in axons. *Nucleic Acids Research*, 46(3):1412–1423, February 2018.

[117] Chong Tang, Rachel Klukovich, Hongying Peng, Zhuqing Wang, Tian Yu, Ying Zhang, Huili Zheng, Arne Klungland, and Wei Yan. ALKBH5-dependent m^6A demethylation controls splicing and stability of long 3′-UTR mRNAs in male germ cells. *Proceedings of the National Academy of Sciences of the United States of America*, 115(2):E325–E333, January 2018.

[118] Wen Xiao, Samir Adhikari, Ujwal Dahal, Yu-Sheng Chen, Ya-Juan Hao, Bao-Fa Sun, Hui-Ying Sun, Ang Li, Xiao-Li Ping, Wei-Yi Lai, Xing Wang, Hai-Li Ma, Chun-Min Huang, Ying Yang, Niu Huang, Gui-Bin Jiang, Hai-Lin Wang, Qi Zhou, Xiu-Jie Wang, Yong-Liang Zhao, and Yun-Gui Yang. Nuclear m^6A Reader YTHDC1 Regulates mRNA Splicing. *Molecular Cell*, 61(4):507–519, February 2016.

[119] Nian Liu, Qing Dai, Guanqun Zheng, Chuan He, Marc Parisien, and Tao Pan. N^6-methyladenosine-dependent RNA structural switches regulate RNA–protein interactions. *Nature*, 518(7540):560–564, February 2015.

[120] Ian A. Roundtree, Guan-Zheng Luo, Zijie Zhang, Xiao Wang, Tao Zhou, Yiquang Cui, Jiahao Sha, Xingxu Huang, Laura Guerrero, Phil Xie, Emily He, Bin Shen, and Chuan He. YTHDC1 mediates nuclear export of N^6-methyladenosine methylated mRNAs. *eLife*, 6, October 2017.

[121] Xiao Wang, Boxuan Simen Zhao, Ian A. Roundtree, Zhike Lu, Dali Han, Honghui Ma, Xiaocheng Weng, Kai Chen, Hailing Shi, and Chuan He. N^6-methyladenosine Modulates Messenger RNA Translation Efficiency. *Cell*, 161(6):1388–1399, June 2015.

[122] Jun Zhou, Ji Wan, Xiangwei Gao, Xingqian Zhang, Samie R. Jaffrey, and Shu-Bing Qian. Dynamic m^6A mRNA methylation directs translational control of heat shock response. *Nature*, 526(7574):591–594, October 2015.

[123] Xiao Wang, Zhike Lu, Adrian Gomez, Gary C. Hon, Yanan Yue, Dali Han, Ye Fu, Marc Parisien, Qing Dai, Guifang Jia, Bing Ren, Tao Pan, and Chuan He. N^6-methyladenosine-dependent regulation of messenger RNA stability. *Nature*, 505(7481):117–120, 2014.

[124] Yamei Niu, Xu Zhao, Yong-Sheng Wu, Ming-Ming Li, Xiu-Jie Wang, and Yun-Gui Yang. N^6-methyl-adenosine (m^6A) in RNA: An Old Modification with A Novel Epigenetic Function. *Genomics, Proteomics & Bioinformatics*, 11(1):8–17, February 2013.

[125] Ye Fu, Gideon Rechavi, and Chuan He. Gene expression regulation mediated through reversible m^{6}A RNA methylation. *Nature Reviews Genetics*, 15(5):293–306, March 2014.

[126] Nian Liu and Tao Pan. N^6-methyladenosine-encoded epitranscriptomics. *Nature Structural & Molecular Biology*, 23(2):98–102, February 2016.

[127] Deepak P. Patil, Brian F. Pickering, and Samie R. Jaffrey. Reading m^{6}A in the Transcriptome: m^{6}A-Binding Proteins. *Trends in Cell Biology*, November 2017.

[128] Dominik Theler, Cyril Dominguez, Markus Blatter, Julien Boudet, and Frédéric H.-T. Allain. Solution structure of the YTH domain in complex with N^6-methyladenosine RNA: a reader of methylated RNA. *Nucleic Acids Research*, 42(22):13911–13919, December 2014.

[129] Shukun Luo and Liang Tong. Molecular basis for the recognition of methylated adenines in RNA by the eukaryotic YTH domain. *Proceedings of the National Academy of Sciences*, 111(38):13834–13839, September 2014.

[130] Chao Xu, Xiao Wang, Ke Liu, Ian A Roundtree, Wolfram Tempel, Yanjun Li, Zhike Lu, Chuan He, and Jinrong Min. Structural basis for selective binding of m^{6}A RNA by the YTHDC1 YTH domain. *Nature Chemical Biology*, 10(11):927–929, September 2014.

[131] Claudio R. Alarcón, Hani Goodarzi, Hyeseung Lee, Xuhang Liu, Saeed Tavazoie, and Sohail F. Tavazoie. HNRNPA2b1 Is a Mediator of m^{6}A-Dependent Nuclear RNA Processing Events. *Cell*, 162(6):1299–1308, September 2015.

[132] Baixing Wu, Shichen Su, Deepak P. Patil, Hehua Liu, Jianhua Gan, Samie R. Jaffrey, and Jinbiao Ma. Molecular Basis For The Specific And Multivariate Recognitions Of RNA Substrates By Human hnRNPA2/B1. *bioRxiv*, page 144345, June 2017.

[133] Stefanie Gerstberger, Markus Hafner, and Thomas Tuschl. A census of human RNA-binding proteins. *Nature Reviews. Genetics*, 15(12):829–845, December 2014.

[134] Hao Du, Ya Zhao, Jinqiu He, Yao Zhang, Hairui Xi, Mofang Liu, Jinbiao Ma, and Ligang Wu. YTHDF2 destabilizes m^{6}A-containing RNA through direct recruitment of the CCR4-NOT deadenylase complex. *Nature Communications*, 7:12626, August 2016.

[135] Ang Li, Yu-Sheng Chen, Xiao-Li Ping, Xin Yang, Wen Xiao, Ying Yang, Hui-Ying Sun, Qin Zhu, Poonam Baidya, Xing Wang, Devi Prasad Bhattarai, Yong-Liang Zhao, Bao-Fa Sun, and Yun-Gui Yang. Cytoplasmic m^{6}A reader YTHDF3 promotes mRNA translation. *Cell Research*, 27(3):444–447, March 2017.

[136] Hailing Shi, Xiao Wang, Zhike Lu, Boxuan S Zhao, Honghui Ma, Phillip J Hsu, and Chuan He. YTHDF3 facilitates translation and decay of N^6-methyladenosine-modified RNA. *Cell Research*, January 2017.

[137] Magdalena Natalia Wojtas, Radha Raman Pandey, Mateusz Mendel, David Homolka, Ravi Sachidanandam, and Ramesh S. Pillai. Regulation of m^6A Transcripts by the $3'$–$5'$ RNA Helicase YTHDC2 Is Essential for a Successful Meiotic Program in the Mammalian Germline. *Molecular Cell*, 68(2):374–387.e12, October 2017.

[138] Kate D. Meyer, Deepak P. Patil, Jun Zhou, Alexandra Zinoviev, Maxim A. Skabkin, Olivier Elemento, Tatyana V. Pestova, Shu-Bing Qian, and Samie R. Jaffrey. $5'$ UTR m^6A Promotes Cap-Independent Translation. *Cell*, 163(4):999–1010, November 2015.

[139] Caroline Roost, Stephen R. Lynch, Pedro J. Batista, Kun Qu, Howard Y. Chang, and Eric T. Kool. Structure and Thermodynamics of N^6-Methyladenosine in RNA: A Spring-Loaded Base Modification. *Journal of the American Chemical Society*, 137(5):2107–2115, February 2015.

[140] Elzbieta Kierzek and Ryszard Kierzek. The thermodynamic stability of RNA duplexes and hairpins containing N^6-alkyladenosines and 2-methylthio-N^6-alkyladenosines. *Nucleic Acids Research*, 31(15):4472–4480, August 2003.

[141] Robert C. Spitale, Ryan A. Flynn, Qiangfeng Cliff Zhang, Pete Crisalli, Byron Lee, Jong-Wha Jung, Hannes Y. Kuchelmeister, Pedro J. Batista, Eduardo A. Torre, Eric T. Kool, and Howard Y. Chang. Structural imprints in vivo decode RNA regulatory mechanisms. *Nature*, 519(7544):486–490, March 2015.

[142] Kathi Zarnack, Julian König, Mojca Tajnik, Iñigo Martincorena, Sebastian Eustermann, Isabelle Stévant, Alejandro Reyes, Simon Anders, Nicholas M. Luscombe, and Jernej Ule. Direct Competition between hnRNP C and U2AF65 Protects the Transcriptome from the Exonization of Alu Elements. *Cell*, 152(3):453–466, January 2013.

[143] Zuzana Cieniková, Fred F. Damberger, Jonathan Hall, Frédéric H.-T. Allain, and Christophe Maris. Structural and Mechanistic Insights into Poly(uridine) Tract Recognition by the hnRNP C RNA Recognition Motif. *Journal of the American Chemical Society*, 136(41):14536–14544, October 2014.

[144] Ravindresh Chhabra. miRNA and methylation: a multifaceted liaison. *Chembiochem: A European Journal of Chemical Biology*, 16(2):195–203, January 2015.

[145] Jinwei Zhang and Robert Landick. A Two-Way Street: Regulatory Interplay between RNA Polymerase and Nascent RNA Structure. *Trends in Biochemical Sciences*, 41(4):293–310, 2016.

[146] Zejuan Li, Hengyou Weng, Rui Su, Xiaocheng Weng, Zhixiang Zuo, Chenying Li, Huilin Huang, Sigrid Nachtergaele, Lei Dong, Chao Hu, Xi Qin, Lichun Tang, Yungui Wang, Gia-Ming Hong, Hao Huang, Xiao Wang, Ping Chen, Sandeep Gurbuxani, Stephen Arnovitz, Yuanyuan Li, Shenglai Li, Jennifer Strong, Mary Beth Neilly, Richard A. Larson, Xi Jiang, Pumin Zhang, Jie Jin, Chuan He, and Jianjun Chen. FTO Plays an Oncogenic Role in Acute Myeloid Leukemia as a N^6-Methyladenosine RNA Demethylase. *Cancer Cell*, 31(1):127–141, January 2017.

[147] Huilin Huang, Hengyou Weng, Wenju Sun, Xi Qin, Hailing Shi, Huizhe Wu, Boxuan Simen Zhao, Ana Mesquita, Chang Liu, Celvie L. Yuan, Yueh-Chiang Hu, Stefan Hüttelmaier, Jennifer R. Skibbe, Rui Su, Xiaolan Deng, Lei Dong, Miao Sun, Chenying Li, Sigrid Nachtergaele, Yungui Wang, Chao Hu, Kyle Ferchen, Kenneth D. Greis, Xi Jiang, Minjie Wei, Lianghu Qu, Jun-Lin Guan, Chuan He, Jianhua Yang, and Jianjun Chen. Recognition of RNA N^6-methyladenosine by IGF2bp proteins enhances mRNA stability and translation. *Nature Cell Biology*, 20(3):285–295, March 2018.

[148] Antoine Cléry, Markus Blatter, and Frédéric H.-T. Allain. RNA recognition motifs: boring? Not quite. *Current Opinion in Structural Biology*, 18(3):290–298, June 2008.

[149] Jean-Yves Roignant and Matthias Soller. m^6A in mRNA: An Ancient Mechanism for Fine-Tuning Gene Expression. *Trends in genetics: TIG*, 33(6):380–390, 2017.

[150] Yang Xiang, Benoit Laurent, Chih-Hung Hsu, Sigrid Nachtergaele, Zhike Lu, Wanqiang Sheng, Chuanyun Xu, Hao Chen, Jian Ouyang, Siqing Wang, Dominic Ling, Pang-Hung Hsu, Lee Zou, Ashwini Jambhekar, Chuan He, and Yang Shi. RNA m^6A methylation regulates the ultraviolet-induced DNA damage response. *Nature*, March 2017.

[151] Boxuan Simen Zhao, Xiao Wang, Alana V. Beadell, Zhike Lu, Hailing Shi, Adam Kuuspalu, Robert K. Ho, and Chuan He. m^6A-dependent maternal mRNA clearance facilitates zebrafish maternal-to-zygotic transition. *Nature*, 542(7642):475–478, 2017.

[152] Shengdong Ke, Endalkachew A. Alemu, Claudia Mertens, Emily Conn Gantman, John J. Fak, Aldo Mele, Bhagwattie Haripal, Ilana Zucker-Scharff, Michael J. Moore, Christopher Y. Park, Cathrine Broberg Vagbø, Anna Kusśnierczyk, Arne Klungland, James E. Darnell, and Robert B. Darnell. A majority of m^6A residues are in the last exons, allowing the potential for $3'$ UTR regulation. *Genes & Development*, 29(19):2037–2053, October 2015.

[153] Claudio R. Alarcón, Hyeseung Lee, Hani Goodarzi, Nils Halberg, and Sohail F. Tavazoie. N^6-methyladenosine marks primary microRNAs for processing. *Nature*, 519(7544):482–485, March 2015.

[154] Annita Louloupi, Evgenia Ntini, Julia Liz, and Ulf Andersson Ørom. Microprocessor dynamics shows co- and post-transcriptional processing of pri-miRNAs. *RNA (New York, N.Y.)*, 23(6):892–898, 2017.

[155] Mariangela Morlando, Monica Ballarino, Natalia Gromak, Francesca Pagano, Irene Bozzoni, and Nick J. Proudfoot. Primary microRNA transcripts are processed cotranscriptionally. *Nature Structural & Molecular Biology*, 15(9):902–909, September 2008.

[156] Shanye Yin, Yong Yu, and Robin Reed. Primary microRNA processing is functionally coupled to RNAP II transcription in vitro. *Scientific Reports*, 5:11992, July 2015.

[157] Matthias W. Hentze, Alfredo Castello, Thomas Schwarzl, and Thomas Preiss. A brave new world of RNA-binding proteins. *Nature Reviews. Molecular Cell Biology*, January 2018.

[158] John Karijolich, Chengqi Yi, and Yi-Tao Yu. Transcriptome-wide dynamics of RNA pseudouridylation. *Nature Reviews. Molecular Cell Biology*, 16(10):581–585, October 2015.

[159] Xiaoyu Li, Shiqing Ma, and Chengqi Yi. Pseudouridine: the fifth RNA nucleotide with renewed interests. *Current Opinion in Chemical Biology*, 33:108–116, August 2016.

[160] Junhui Ge and Yi-Tao Yu. RNA pseudouridylation: new insights into an old modification. *Trends in Biochemical Sciences*, 38(4):210–218, April 2013.

[161] Katherine E. Sloan, Ahmed S. Warda, Sunny Sharma, Karl-Dieter Entian, Denis L. J. Lafontaine, and Markus T. Bohnsack. Tuning the ribosome: The influence of rRNA modification on eukaryotic ribosome biogenesis and function. *RNA Biology*, 14(9):1138–1152, September 2017.

[162] Tomasz W. Turowski and David Tollervey. Cotranscriptional events in eukaryotic ribosome synthesis. *Wiley interdisciplinary reviews. RNA*, 6(1):129–139, February 2015.

[163] Meredith I. Newby and Nancy L. Greenbaum. Investigation of Overhauser effects between pseudouridine and water protons in RNA helices. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12697–12702, October 2002.

[164] Graham A. Hudson, Richard J. Bloomingdale, and Brent M. Znosko. Thermodynamic contribution and nearest-neighbor parameters of pseudouridine-adenosine base pairs in oligoribonucleotides. *RNA (New York, N.Y.)*, 19(11):1474–1482, November 2013.

[165] Elzbieta Kierzek, Magdalena Malgowska, Jolanta Lisowiec, Douglas H. Turner, Zofia Gdaniec, and Ryszard Kierzek. The contribution of pseudouridine to stabilities and structure of RNAs. *Nucleic Acids Research*, 42(5):3492–3501, March 2014.

[166] Thomas H. King, Ben Liu, Ryan R. McCully, and Maurille J. Fournier. Ribosome structure and activity are altered in cells lacking snoRNPs that form pseudouridines in the peptidyl transferase center. *Molecular Cell*, 11(2):425–435, February 2003.

[167] Meredith I. Newby and Nancy L. Greenbaum. Sculpting of the spliceosomal branch site recognition motif by a conserved pseudouridine. *Nature Structural Biology*, 9(12):958–965, December 2002.

[168] Karen Jack, Cristian Bellodi, Dori M. Landry, Rachel O. Niederer, Arturas Meskauskas, Sharmishtha Musalgaonkar, Noam Kopmar, Olya Krasnykh, Alison M. Dean, Sunnie R. Thompson, Davide Ruggero, and Jonathan D. Dinman. rRNA pseudouridylation defects affect ribosomal ligand binding and translational fidelity from yeast to human cells. *Molecular Cell*, 44(4):660–666, November 2011.

[169] John Karijolich and Yi-Tao Yu. Converting nonsense codons into sense codons by targeted pseudouridylation. *Nature*, 474(7351):395–398, June 2011.

[170] Marc Parisien, Chengqi Yi, and Tao Pan. Rationalization and prediction of selective decoding of pseudouridine-modified nonsense and sense codons. *RNA (New York, N.Y.)*, 18(3):355–367, March 2012.

[171] Stefanie A. Mortimer, Mary Anne Kidwell, and Jennifer A. Doudna. Insights into RNA structure and function from genome-wide studies. *Nature Reviews. Genetics*, 15(7):469–479, July 2014.

[172] Masato Taoka, Yuko Nobe, Yuka Yamaki, Yoshio Yamauchi, Hideaki Ishikawa, Nobuhiro Takahashi, Hiroshi Nakayama, and Toshiaki Isobe. The complete chemical structure of Saccharomyces cerevisiae rRNA: partial pseudouridylation of U2345 in 25s rRNA by snoRNA snR9. *Nucleic Acids Research*, 44(18):8951–8961, October 2016.

[173] Guowei Wu, Mu Xiao, Chunxing Yang, and Yi-Tao Yu. U2 snRNA is inducibly pseudouridylated at novel sites by Pus7p and snR81 RNP. *The EMBO journal*, 30(1):79–89, January 2011.

[174] Boxuan Simen Zhao and Chuan He. Pseudouridine in a new era of RNA modifications. *Cell Research*, 25(2):153–154, February 2015.

[175] Pavanapuresan P. Vaidyanathan, Ishraq AlSadhan, Dawn K. Merriman, Hashim M. Al-Hashimi, and Daniel Herschlag. Pseudouridine and N^6-methyladenosine modifications weaken PUF protein/RNA interactions. *RNA (New York, N.Y.)*, 23(5):611–618, 2017.

[176] A. Emilia Arguello, Amanda N. DeLiberto, and Ralph E. Kleiner. RNA Chemical Proteomics Reveals the N^6-Methyladenosine (m^6A)-Regulated Protein–RNA Interactome. *Journal of the American Chemical Society*, 139(48):17249–17252, December 2017.

[177] Raghu R. Edupuganti, Simon Geiger, Rik G. H. Lindeboom, Hailing Shi, Phillip J. Hsu, Zhike Lu, Shuang-Yin Wang, Marijke P. A. Baltissen, Pascal W. T. C. Jansen, Martin Rossa, Markus Müller, Hendrik G. Stunnenberg, Chuan He, Thomas Carell, and Michiel Vermeulen. N^6-methyladenosine (m^6A) recruits and repels proteins to regulate mRNA homeostasis. *Nature Structural & Molecular Biology*, 24(10):870–878, October 2017.

[178] Andrey Bakin and James Ofengand. Four newly located pseudouridylate residues in Escherichia coli 23S ribosomal RNA are all at the peptidyltransferase center: Analysis by the application of a new sequencing technique. *Biochemistry*, 32(37):9754–9762, September 1993.

[179] Xiaoyu Li, Ping Zhu, Shiqing Ma, Jinghui Song, Jinyi Bai, Fangfang Sun, and Chengqi Yi. Chemical pulldown reveals dynamic pseudouridylation of the mammalian transcriptome. *Nature Chemical Biology*, 11(8):592–597, August 2015.

[180] Carmelita Nora Marbaniang and Jörg Vogel. Emerging roles of RNA modifications in bacteria. *Current Opinion in Microbiology*, 30:50–57, April 2016.

[181] Wendy V. Gilbert, Tristan A. Bell, and Cassandra Schaening. Messenger RNA modifications: Form, distribution, and function. *Science*, 352(6292):1408–1412, 2016.

[182] Schraga Schwartz and Yuri Motorin. Next-generation sequencing technologies for detection of modified nucleotides in RNAs. *RNA biology*, 14(9):1124–1137, September 2017.

[183] Jacob C. Schwartz, Xueyin Wang, Elaine R. Podell, and Thomas R. Cech. RNA Seeds Higher-Order Assembly of FUS Protein. *Cell Reports*, 5(4):918–925, November 2013.

[184] Maryam Zaringhalam and F. Nina Papavasiliou. Pseudouridylation meets next-generation sequencing. *Methods (San Diego, Calif.)*, 107:63–72, September 2016.

[185] Kai Chen, Zhike Lu, Xiao Wang, Ye Fu, Guan-Zheng Luo, Nian Liu, Dali Han, Dan Dominissini, Qing Dai, Tao Pan, and Chuan He. High-Resolution N^6-Methyladenosine (m^6A) Map Using Photo-Crosslinking-Assisted m^6A Sequencing. *Angewandte Chemie International Edition*, 54(5):1587–1590, 2015.

[186] Benoit Molinie, Jinkai Wang, Kok Seong Lim, Roman Hillebrand, Zhi-Xiang Lu, Nicholas Van Wittenberghe, Benjamin D. Howard, Kaveh Daneshvar, Alan C. Mullen, Peter Dedon, Yi Xing, and Cosmas C. Giallourakis. m^6A-LAIC-seq reveals the census and complexity of the m^6A epitranscriptome. *Nature Methods*, 13(8):692–698, 2016.

[187] Heidelinde Glasner, Christian Riml, Ronald Micura, and Kathrin Breuker. Label-free, direct localization and relative quantitation of the RNA nucleobase methylations m^6A, m^5C, m^3U, and m^5U by top-down mass spectrometry. *Nucleic Acids Research*, 45(13):8014–8025, July 2017.

[188] Ralf Hauenschild, Lyudmil Tserovski, Katharina Schmid, Kathrin Thüring, Marie-Luise Winz, Sunny Sharma, Karl-Dieter Entian, Ludivine Wacheul, Denis L. J. Lafontaine, James Anderson, Juan Alfonzo, Andreas Hildebrandt, Andres Jäschke, Yuri Motorin, and Mark Helm. The reverse transcription signature of N^1-methyladenosine in RNA-Seq is sequence dependent. *Nucleic Acids Research*, 43(20):9950–9964, November 2015.

[189] Ralf Hauenschild, Stephan Werner, Lyudmil Tserovski, Andreas Hildebrandt, Yuri Motorin, and Mark Helm. CoverageAnalyzer (CAn): A Tool for Inspection of Modification Signatures in RNA Sequencing Profiles. *Biomolecules*, 6(4), 2016.

[190] Mark Helm and Yuri Motorin. Detecting RNA modifications in the epitranscriptome: predict and validate. *Nature Reviews. Genetics*, 18(5):275–291, 2017.

[191] Bradley M. Lunde, Claire Moore, and Gabriele Varani. RNA-binding proteins: modular design for efficient function. *Nature Reviews Molecular Cell Biology*, 8(6):479–490, June 2007.

[192] Yaseswini Neelamraju, Seyedsasan Hashemikhabir, and Sarath Chandra Janga. The human RBPome: From genes and proteins to human disease. *Journal of Proteomics*, 127, Part A:61–70, September 2015.

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

[194] Roberto Valverde, Laura Edwards, and Lynne Regan. Structure and function of KH domains. *The FEBS journal*, 275(11):2712–2726, June 2008.

[195] Grégoire Masliah, Pierre Barraud, and Frédéric H.-T. Allain. RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence. *Cellular and molecular life sciences: CMLS*, 70(11):1875–1895, June 2013.

[196] Traci M. Tanaka Hall. Multiple modes of RNA recognition by zinc finger proteins. *Current Opinion in Structural Biology*, 15(3):367–373, June 2005.

[197] William H. Hudson and Eric A. Ortlund. The structure, function and evolution of proteins that bind DNA and RNA. *Nature Reviews. Molecular Cell Biology*, 15(11):749–760, November 2014.

[198] Alfredo Castello, Bernd Fischer, Katrin Eichelbaum, Rastislav Horos, Benedikt M. Beckmann, Claudia Strein, Norman E. Davey, David T. Humphreys, Thomas Preiss, Lars M. Steinmetz, Jeroen Krijgsveld, and Matthias W. Hentze. Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell*, 149(6):1393–1406, June 2012.

[199] Robin van der Lee, Marija Buljan, Benjamin Lang, Robert J. Weatheritt, Gary W. Daughdrill, A. Keith Dunker, Monika Fuxreiter, Julian Gough, Joerg Gsponer, David T. Jones, Philip M. Kim, Richard W. Kriwacki, Christopher J. Oldfield, Rohit V. Pappu, Peter Tompa, Vladimir N. Uversky, Peter E. Wright, and M. Madan Babu. Classification of intrinsically disordered regions and proteins. *Chemical Reviews*, 114(13):6589–6631, July 2014.

[200] Alain Coletta, John W. Pinney, David Y. Weiss Solís, James Marsh, Steve R. Pettifer, and Teresa K. Attwood. Low-complexity regions within protein sequences have position-dependent roles. *BMC systems biology*, 4:43, April 2010.

[201] Mark A. DePristo, Martine M. Zilversmit, and Daniel L. Hartl. On the abundance, amino acid composition, and evolutionary dynamics of low-complexity regions in proteins. *Gene*, 378:19–30, August 2006.

[202] Bandana Kumari, Ravindra Kumar, and Manish Kumar. Low complexity and disordered regions of proteins have different structural and amino acid preferences. *Molecular bioSystems*, 11(2):585–594, February 2015.

[203] Benedikt M. Beckmann, Alfredo Castello, and Jan Medenbach. The expanding universe of ribonucleoproteins: of novel RNA-binding proteins and unconventional interactions. *Pflugers Archiv: European Journal of Physiology*, 468(6):1029–1040, 2016.

[204] Aino I. Järvelin, Marko Noerenberg, Ilan Davis, and Alfredo Castello. The new (dis)order in RNA regulation. *Cell communication and signaling: CCS*, 14:9, April 2016.

[205] Susan M. Corley and Jill E. Gready. Identification of the RGG box motif in Shadoo: RNA-binding and signaling roles? *Bioinformatics and Biology Insights*, 2:383–400, November 2008.

[206] Palaniraja Thandapani, Timothy R. O'Connor, Timothy L. Bailey, and Stéphane Richard. Defining the RGG/RG Motif. *Molecular Cell*, 50(5):613–623, June 2013.

[207] Eva Schöller, Franziska Weichmann, Thomas Treiber, Sam Ringle, Nora Treiber, Andrew Flatley, Regina Feederle, Astrid Bruckmann, and Gunter Meister. Interactions, localization, and phosphorylation of the m^6A generating METTL3–METTL14–WTAP complex. *RNA (New York, N.Y.)*, 24(4):499–512, April 2018.

[208] Anh Tuân Phan, Vitaly Kuryavyi, Jennifer C. Darnell, Alexander Serganov, Ananya Majumdar, Serge Ilin, Tanya Raslin, Anna Polonskaia, Cynthia Chen, David Clain, Robert B. Darnell, and Dinshaw J. Patel. Structure-function studies of FMRP RGG peptide recognition of an RNA duplex-quadruplex junction. *Nature Structural & Molecular Biology*, 18(7):796–804, June 2011.

[209] Nikita Vasilyev, Anna Polonskaia, Jennifer C. Darnell, Robert B. Darnell, Dinshaw J. Patel, and Alexander Serganov. Crystal structure reveals specific recognition of a G-quadruplex RNA by a β-turn in the RGG motif of FMRP. *Proceedings of the National Academy of Sciences of the United States of America*, 112(39):E5391–5400, September 2015.

[210] Bagdeser A. Ozdilek, Valery F. Thompson, Nasiha S. Ahmed, Connor I. White, Robert T. Batey, and Jacob C. Schwartz. Intrinsically disordered RGG/RG domains mediate degenerate specificity in RNA binding. *Nucleic Acids Research*, 45(13):7984–7996, July 2017.

[211] Purusharth Rajyaguru, Meipei She, and Roy Parker. Scd6 targets eIF4g to repress translation: RGG motif proteins as a class of eIF4g-binding proteins. *Molecular Cell*, 45(2):244–254, January 2012.

[212] Ambrosius P. Snijders, Guillaume M. Hautbergue, Alex Bloom, James C. Williamson, Thomas C. Minshull, Helen L. Phillips, Simeon R. Mihaylov, Douglas T. Gjerde, David P. Hornby, Stuart A. Wilson, Paul J. Hurd, and Mark J. Dickman. Arginine methylation and citrullination of splicing factor proline- and glutamine-rich (SFPQ/PSF) regulates its association with mRNA. *RNA (New York, N.Y.)*, 21(3):347–359, March 2015.

[213] Yanzhong Yang and Mark T. Bedford. Protein arginine methyltransferases and cancer. *Nature Reviews. Cancer*, 13(1):37–50, January 2013.

[214] Vladimir N. Uversky. Intrinsically disordered proteins in overcrowded milieu: Membrane-less organelles, phase separation, and intrinsic disorder. *Current Opinion in Structural Biology*, 44:18–30, 2017.

[215] Edward W. J. Wallace, Jamie L. Kear-Scott, Evgeny V. Pilipenko, Michael H. Schwartz, Pawel R. Laskowski, Alexandra E. Rojek, Christopher D. Katanski, Joshua A. Riback, Michael F. Dion, Alexander M. Franks, Edoardo M. Airoldi, Tao Pan, Bogdan A. Budnik, and D. Allan Drummond. Reversible, Specific, Active Aggregates of Endogenous Proteins Assemble upon Heat Stress. *Cell*, 162(6):1286–1298, September 2015.

[216] Briana Van Treeck, David S. W. Protter, Tyler Matheny, Anthony Khong, Christopher D. Link, and Roy Parker. RNA self-assembly contributes to stress granule formation and defining the stress granule transcriptome. *Proceedings of the National Academy of Sciences of the United States of America*, 115(11):2734–2739, March 2018.

[217] Sonja Kroschwald, Shovamayee Maharana, Daniel Mateju, Liliana Malinovska, Elisabeth Nüske, Ina Poser, Doris Richter, and Simon Alberti. Promiscuous interactions and protein disaggregases determine the material state of stress-inducible RNP granules. *eLife*, 4:e06807, August 2015.

[218] Clifford P. Brangwynne, Christian R. Eckmann, David S. Courson, Agata Rybarska, Carsten Hoege, Jöbin Gharakhani, Frank Jülicher, and Anthony A. Hyman. Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science (New York, N.Y.)*, 324(5935):1729–1732, June 2009.

[219] Amandine Molliex, Jamshid Temirov, Jihun Lee, Maura Coughlin, Anderson P. Kanagaraj, Hong Joo Kim, Tanja Mittag, and J. Paul Taylor. Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. *Cell*, 163(1):123–133, September 2015.

[220] Joshua R. Wheeler, Tyler Matheny, Saumya Jain, Robert Abrisch, and Roy Parker. Distinct stages in stress granule assembly and disassembly. *eLife*, 5, 2016.

[221] Ankur Jain and Ronald D. Vale. RNA phase transitions in repeat expansion disorders. *Nature*, 546(7657):243–247, 2017.

[222] Joshua A. Riback, Christopher D. Katanski, Jamie L. Kear-Scott, Evgeny V. Pilipenko, Alexandra E. Rojek, Tobin R. Sosnick, and D. Allan Drummond. Stress-Triggered Phase Separation Is an Adaptive, Evolutionarily Tuned Response. *Cell*, 168(6):1028–1040.e19, March 2017.

[223] Tina W. Han, Masato Kato, Shanhai Xie, Leeju C. Wu, Hamid Mirzaei, Jimin Pei, Min Chen, Yang Xie, Jeffrey Allen, Guanghua Xiao, and Steven L. McKnight. Cell-free Formation of RNA Granules: Bound RNAs Identify Features and Components of Cellular Assemblies. *Cell*, 149(4):768–779, May 2012.

[224] Masato Kato, Tina W. Han, Shanhai Xie, Kevin Shi, Xinlin Du, Leeju C. Wu, Hamid Mirzaei, Elizabeth J. Goldsmith, Jamie Longgood, Jimin Pei, Nick V. Grishin, Douglas E. Frantz, Jay W. Schneider, She Chen, Lin Li, Michael R. Sawaya, David Eisenberg, Robert Tycko, and Steven L. McKnight. Cell-free Formation of RNA Granules: Low Complexity Sequence Domains Form Dynamic Fibers within Hydrogels. *Cell*, 149(4):753–767, May 2012.

[225] Kathleen A. Burke, Abigail M. Janke, Christy L. Rhine, and Nicolas L. Fawzi. Residue-by-Residue View of In Vitro FUS Granules that Bind the C-Terminal Domain of RNA Polymerase II. *Molecular Cell*, 60(2):231–241, October 2015.

[226] Yuan Lin, David S. W. Protter, Michael K. Rosen, and Roy Parker. Formation and Maturation of Phase-Separated Liquid Droplets by RNA-Binding Proteins. *Molecular Cell*, 60(2):208–219, October 2015.

[227] Huaiying Zhang, Shana Elbaum-Garfinkle, Erin M. Langdon, Nicole Taylor, Patricia Occhipinti, Andrew A. Bridges, Clifford P. Brangwynne, and Amy S. Gladfelter. RNA Controls PolyQ Protein Phase Transitions. *Molecular Cell*, 60(2):220–230, October 2015.

[228] Adam G. Larson, Daniel Elnatan, Madeline M. Keenen, Michael J. Trnka, Jonathan B. Johnston, Alma L. Burlingame, David A. Agard, Sy Redding, and Geeta J. Narlikar. Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin. *Nature*, 547(7662):236–240, 2017.

[229] Amy R. Strom, Alexander V. Emelyanov, Mustafa Mir, Dmitry V. Fyodorov, Xavier Darzacq, and Gary H. Karpen. Phase separation drives heterochromatin domain formation. *Nature*, 547(7662):241–245, 2017.

[230] Ilmin Kwon, Masato Kato, Siheng Xiang, Leeju Wu, Pano Theodoropoulos, Hamid Mirzaei, Tina Han, Shanhai Xie, Jeffry L. Corden, and Steven L. McKnight. Phosphorylation-Regulated Binding of RNA Polymerase II to Fibrous Polymers of Low-Complexity Domains. *Cell*, 155(5):1049–1060, November 2013.

[231] Denes Hnisz, Krishna Shrinivas, Richard A. Young, Arup K. Chakraborty, and Phillip A. Sharp. A Phase Separation Model for Transcriptional Control. *Cell*, 169(1):13–23, March 2017.

[232] Yue Wan, Michael Kertesz, Robert C. Spitale, Eran Segal, and Howard Y. Chang. Understanding the transcriptome through RNA structure. *Nature Reviews Genetics*, 12(9):641–655, August 2011.

[233] Kévin Darty, Alain Denise, and Yann Ponty. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, 25(15):1974–1975, August 2009.

[234] Marc Parisien and François Major. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, 452(7183):51–55, March 2008.

[235] Michael P. Latham, Darin J. Brown, Scott A. McCallum, and Arthur Pardi. NMR Methods for Studying the Structure and Dynamics of RNA. *ChemBioChem*, 6(9):1492–1505, September 2005.

[236] Lukáš Zídek, Richard Štefl, and Vladimíř Sklenár. NMR methodology for the study of nucleic acids. *Current Opinion in Structural Biology*, 11(3):275–281, June 2001.

[237] Shimon Weiss. Measuring conformational dynamics of biomolecules by single molecule fluorescence spectroscopy. *Nature Structural & Molecular Biology*, 7(9):724–729, September 2000.

[238] Elvin A Alemlán, Rajan Lamichhane, and David Rueda. Exploring RNA folding one molecule at a time. *Current Opinion in Chemical Biology*, 12(6):647–654, December 2008.

[239] Qing Dai, Robert Fong, Mridusmita Saikia, David Stephenson, Yi-tao Yu, Tao Pan, and Joseph A. Piccirilli. Identification of recognition residues for ligation-based detection and quantitation of pseudouridine and N^6-methyladenosine. *Nucleic Acids Research*, 35(18):6322–6329, August 2007.

[240] Marc Parisien, José A. Cruz, Éric Westhof, and François Major. New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA*, 15(10):1875–1885, October 2009.

[241] Bettina Heinrich, Zhaiyi Zhang, Oleg Raitskin, Michael Hiller, Natalya Benderska, Annette M. Hartmann, Laurent Bracco, David Elliott, Shani Ben-Ari, Hermona Soreq, Joseph Sperling, Ruth Sperling, and Stefan Stamm. Heterogeneous Nuclear Ribonucleoprotein G Regulates Splice Site Selection by Binding to CC(A/C)-rich Regions in Pre-mRNA. *Journal of Biological Chemistry*, 284(21):14303–14315, May 2009.

[242] Yvonne Hofmann and Brunhilde Wirth. hnRNP-G promotes exon 7 inclusion of survival motor neuron (SMN) via direct interaction with Htra2-1. *Human molecular genetics*, 11(17):2037–2049, 2002.

[243] Yan Wang, Junning Wang, Lei Gao, Stefan Stamm, and Athena Andreadis. An SRp75/hnRNPG complex interacting with hnRNPE2 regulates the $5'$ splice site of tau exon 10, whose misregulation causes frontotemporal dementia. *Gene*, 485(2):130–138, October 2011.

[244] Britt Adamson, Agata Smogorzewska, Frederic D. Sigoillot, Randall W. King, and Stephen J. Elledge. A genome-wide homologous recombination screen identifies the RNA-binding protein RBMX as a component of the DNA-damage response. *Nature Cell Biology*, 14(3):318–328, February 2012.

[245] Sachihiro Matsunaga, Hideaki Takata, Akihiro Morimoto, Kayoko Hayashihara, Tsunehito Higashi, Kouhei Akatsuchi, Eri Mizusawa, Mariko Yamakawa, Mamoru Ashida, Tomoko M. Matsunaga, Takachika Azuma, Susumu Uchiyama, and Kiichi Fukui. RBMX: A Regulator for Maintenance and Centromeric Protection of Sister Chromatid Cohesion. *Cell Reports*, 1(4):299–308, April 2012.

[246] Sheng Zhao, Wayne J. Korzan, Chun-Chun Chen, and Russell D. Fernald. Heterogeneous nuclear ribonucleoprotein A/B and G inhibits the transcription of gonadotropin-releasing-hormone 1. *Molecular and Cellular Neuroscience*, 37(1):69–84, January 2008.

[247] Enkhjargal Tsend-Ayush, Lynda A. O'Sullivan, Frank S. Grützner, Sara M.N. Onnebo, Rowena S. Lewis, Margaret L. Delbridge, Jennifer A. Marshall Graves, and Alister C. Ward. *RBMX* gene is essential for brain development in zebrafish. *Developmental Dynamics*, 234(3):682–688, November 2005.

[248] V. Shashi, P. Xie, K. Schoch, D.B. Goldstein, T.D. Howard, M.N. Berry, C.E. Schwartz, K. Cronin, S. Sliwa, A. Allen, and A.C. Need. The *RBMX* gene as a candidate for the Shashi X-linked intellectual disability syndrome: *RBMX* gene for SMRXS. *Clinical Genetics*, 88(4):386–390, October 2015.

[249] Rasha Kanhoush, Brent Beenders, Caroline Perrin, Jacques Moreau, Michel Bellini, and May Penrad-Mobayed. Novel domains in the hnRNP G/RBMX protein with distinct roles in RNA binding and targeting nascent transcripts. *Nucleus*, 1(1):109–122, 2010.

[250] Yue Wan, Kun Qu, Qiangfeng Cliff Zhang, Ryan A. Flynn, Ohad Manor, Zhengqing Ouyang, Jiajing Zhang, Robert C. Spitale, Michael P. Snyder, Eran Segal, and Howard Y. Chang. Landscape and variation of RNA secondary structure across the human transcriptome. *Nature*, 505(7485):706–709, January 2014.

[251] Cole Trapnell, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R. Kelley, Harold Pimentel, Steven L. Salzberg, John L. Rinn, and Lior Pachter. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7(3):562–578, March 2012.

[252] Simon Anders, Alejandro Reyes, and Wolfgang Huber. Detecting differential usage of exons from RNA-seq data. *Genome Research*, 22(10):2008–2017, October 2012.

[253] Mani Ramaswami, J. Paul Taylor, and Roy Parker. Altered Ribostasis: RNA-Protein Granules in Degenerative Disorders. *Cell*, 154(4):727–736, August 2013.

[254] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano, Anna-Carina Jungkamp, Mathias Munschauer, Alexander Ulrich, Greg S. Wardle, Scott Dewell, Mihaela Zavolan, and Thomas Tuschl. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1):129–141, April 2010.

[255] David L. Corcoran, Stoyan Georgiev, Neelanjan Mukherjee, Eva Gottwein, Rebecca L. Skalsky, Jack D. Keene, and Uwe Ohler. PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. *Genome Biology*, 12(8):R79, 2011.

[256] Dan Dominissini, Sharon Moshitch-Moshkovitz, Mali Salmon-Divon, Ninette Amariglio, and Gideon Rechavi. Transcriptome-wide mapping of N^6-methyladenosine by m^{6}A-seq based on immunocapturing and massively parallel sequencing. *Nature Protocols*, 8(1):176–189, January 2013.

[257] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.

[258] Cole Trapnell, Lior Pachter, and Steven L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, May 2009.

[259] Barbara N. Borsos, Ildikó Huliák, Hajnalka Majoros, Zsuzsanna Ujfaludi, Ákos Gyenis, Peter Pukler, Imre M. Boros, and Tibor Pankotai. Human p53 interacts with the elongating RNAPII complex and is required for the release of actinomycin D induced transcription blockage. *Scientific Reports*, 7:40960, January 2017.

[260] Olivier Sordet, Stéphane Larochelle, Estelle Nicolas, Ellen V. Stevens, Chao Zhang, Kevan M. Shokat, Robert P. Fisher, and Yves Pommier. Hyperphosphorylation of RNA polymerase II in response to topoisomerase I cleavage complexes and its association with transcription- and BRCA1-dependent degradation of topoisomerase I. *Journal of Molecular Biology*, 381(3):540–549, September 2008.

[261] Ahmed Moursy, Frédéric H.-T. Allain, and Antoine Cléry. Characterization of the RNA recognition mode of hnRNP G extends its role in SMN2 splicing regulation. *Nucleic Acids Research*, 42(10):6659–6672, June 2014.

[262] Zsuzsanna Dosztányi, Veronika Csizmók, Peter Tompa, and István Simon. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *Journal of Molecular Biology*, 347(4):827–839, April 2005.

[263] Zsuzsanna Dosztányi, Veronika Csizmók, Peter Tompa, and István Simon. IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics (Oxford, England)*, 21(16):3433–3434, August 2005.

[264] Darwin S. Dichmann, Russell B. Fletcher, and Richard M. Harland. Expression cloning in Xenopus identifies RNA-binding proteins as regulators of embryogenesis and Rbmx as necessary for neural and muscle development. *Developmental Dynamics: An Official Publication of the American Association of Anatomists*, 237(7):1755–1766, July 2008.

[265] M. Talat Nasim, Tatyana K. Chernova, Hasnin M. Chowdhury, Bai-Gong Yue, and Ian C. Eperon. HnRNP G and Tra2beta: opposite effects on splicing matched by antagonism in RNA binding. *Human Molecular Genetics*, 12(11):1337–1348, June 2003.

[266] Kresten Lindorff-Larsen, Stefano Piana, Ron O. Dror, and David E. Shaw. How fast-folding proteins fold. *Science (New York, N.Y.)*, 334(6055):517–520, October 2011.

[267] John J. Skinner, Wookyung Yu, Elizabeth K. Gichana, Michael C. Baxa, James R. Hinshaw, Karl F. Freed, and Tobin R. Sosnick. Benchmarking all-atom simulations using hydrogen exchange. *Proceedings of the National Academy of Sciences of the United States of America*, 111(45):15975–15980, November 2014.

[268] Demet Araç, Antony A. Boucard, Marc F. Bolliger, Jenna Nguyen, S. Michael Soltis, Thomas C. Südhof, and Axel T. Brunger. A novel evolutionarily conserved domain of cell-adhesion GPCRs mediates autoproteolysis. *The EMBO journal*, 31(6):1364–1378, March 2012.

[269] Christina M. Ferrer, Marielle Alders, Alex V. Postma, Seonmi Park, Mark A. Klein, Murat Cetinbas, Eva Pajkrt, Astrid Glas, Silvana van Koningsbruggen, Vincent M. Christoffels, Marcel M. A. M. Mannens, Lia Knegt, Jean-Pierre Etchegaray, Ruslan I. Sadreyev, John M. Denu, Gustavo Mostoslavsky, Merel C. van Maarle, and Raul Mostoslavsky. An inactivating mutation in the histone deacetylase SIRT6 causes human perinatal lethality. *Genes & Development*, 32(5-6):373–388, March 2018.

[270] Regina Groisman, Jolanta Polanowska, Isao Kuraoka, Jun-ichi Sawada, Masafumi Saijo, Ronny Drapkin, Alexei F. Kisselev, Kiyoji Tanaka, and Yoshihiro Nakatani. The ubiquitin ligase activity in the DDB2 and CSA complexes is differentially regulated by the COP9 signalosome in response to DNA damage. *Cell*, 113(3):357–367, May 2003.

[271] Masahiro Okada and Tatsuo Fukagawa. Purification of a protein complex that associates with chromatin. *Protocol Exchange*, DOI: 10.1038/nprot.2006.417, December 2006.

[272] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England)*, 29(1):15–21, January 2013.

[273] Nuala A. O'Leary, Mathew W. Wright, J. Rodney Brister, Stacy Ciufo, Diana Haddad, Rich McVeigh, Bhanu Rajput, Barbara Robbertse, Brian Smith-White, Danso Ako-Adjei, Alexander Astashyn, Azat Badretdin, Yiming Bao, Olga Blinkova, Vyacheslav Brover, Vyacheslav Chetvernin, Jinna Choi, Eric Cox, Olga Ermolaeva, Catherine M. Farrell, Tamara Goldfarb, Tripti Gupta, Daniel Haft, Eneida Hatcher, Wratko Hlavina, Vinita S. Joardar, Vamsi K. Kodali, Wenjun Li, Donna Maglott, Patrick Masterson, Kelly M. McGarvey, Michael R. Murphy, Kathleen O'Neill, Shashikant Pujar, Sanjida H. Rangwala, Daniel Rausch, Lillian D. Riddick, Conrad Schoch, Andrei Shkeda, Susan S. Storz, Hanzhen Sun, Francoise Thibaud-Nissen, Igor Tolstoy, Raymond E. Tully, Anjana R. Vatsan, Craig Wallin, David Webb, Wendy Wu, Melissa J. Landrum, Avi Kimchi, Tatiana Tatusova, Michael DiCuccio, Paul Kitts, Terence D. Murphy, and Kim D. Pruitt. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, 44(D1):D733–745, January 2016.

[274] Donna Karolchik, Angela S. Hinrichs, Terrence S. Furey, Krishna M. Roskin, Charles W. Sugnet, David Haussler, and W. James Kent. The UCSC Table Browser data retrieval tool. *Nucleic Acids Research*, 32(Database issue):D493–496, January 2004.

[275] Yang Liao, Gordon K. Smyth, and Wei Shi. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics (Oxford, England)*, 30(7):923–930, April 2014.

[276] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014.

[277] Yufei Xiao, Tzu-Hung Hsiao, Uthra Suresh, Hung-I. Harry Chen, Xiaowu Wu, Steven E. Wolf, and Yidong Chen. A novel significance score for gene selection and ranking. *Bioinformatics (Oxford, England)*, 30(6):801–807, March 2014.

[278] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, 25(1):25–29, May 2000.

[279] The Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research*, 45(D1):D331–D338, January 2017.

[280] Huaiyu Mi, Xiaosong Huang, Anushya Muruganujan, Haiming Tang, Caitlin Mills, Diane Kang, and Paul D. Thomas. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Research*, 45(D1):D183–D189, January 2017.

[281] Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6):841–842, March 2010.

[282] Nathan A. Siegfried, Steven Busan, Greggory M. Rice, Julie A. E. Nelson, and Kevin M. Weeks. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). *Nature Methods*, 11(9):959–965, September 2014.

[283] Matthew J. Smola, Greggory M. Rice, Steven Busan, Nathan A. Siegfried, and Kevin M. Weeks. Selective $2'$-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. *Nature Protocols*, 10(11):1643–1669, November 2015.

[284] Alessandro Calabretta and Christian J. Leumann. Base pairing and miscoding properties of $1,N^6$-ethenoadenine- and $3,N^4$-ethenocytosine-containing RNA oligonucleotides. *Biochemistry*, 52(11):1990–1997, March 2013.

[285] Dorota Piekna-Przybylska, Wayne A. Decatur, and Maurille J. Fournier. The 3D rRNA modification maps database: with interactive tools for ribosome analysis. *Nucleic Acids Research*, 36(Database issue):D178–183, January 2008.

[286] Jaclyn A. Miranda and Grieg F. Steward. Variables influencing the efficiency and interpretation of reverse transcription quantitative PCR (RT-qPCR): An empirical study using Bacteriophage MS2. *Journal of Virological Methods*, 241:1–10, 2017.

[287] Anders Stahlberg, Joakim Hakansson, Xiaojie Xian, Henrik Semb, and Mikael Kubista. Properties of the Reverse Transcription Reaction in mRNA Quantification. *Clinical Chemistry*, 50(3):509–515, March 2004.

[288] Carole Bampi, Arkadiusz Bibillo, Michaela Wendeler, Gilles Divita, Robert J. Gorelick, Stuart F. J. Le Grice, and Jean-Luc Darlix. Nucleotide Excision Repair and Template-independent Addition by HIV-1 Reverse Transcriptase in the Presence of Nucleocapsid Protein. *Journal of Biological Chemistry*, 281(17):11736–11743, April 2006.

[289] Yuegao Huang, Congju Chen, and Irina M. Russu. Dynamics and Stability of Individual Base Pairs in Two Homologous RNA-DNA Hybrids. *Biochemistry*, 48(18):3988–3997, May 2009.

[290] Pawel Dabrowski-Tumanski, Joanna Kowalska, and Jacek Jemielity. Efficient and Rapid Synthesis of Nucleoside Diphosphate Sugars from Nucleoside Phosphorimidazolides. *European Journal of Organic Chemistry*, 2013(11):2147–2154, April 2013.

[291] Tianlei Li, Abdellatif Tikad, Weidong Pan, and Stéphane P. Vincent. -Stereoselective phosphorylations applied to the synthesis of ADP- and polyprenyl--mannopyranosides. *Organic Letters*, 16(21):5628–5631, November 2014.

[292] Wesley C. Clark, Molly E. Evans, Dan Dominissini, Guanqun Zheng, and Tao Pan. tRNA base methylation identification and quantification via high-throughput sequencing. *RNA (New York, N.Y.)*, 22(11):1771–1784, November 2016.

[293] Gianluigi Lichinchi, Shang Gao, Yogesh Saletore, Gwendolyn Michelle Gonzalez, Vikas Bansal, Yinsheng Wang, Christopher E. Mason, and Tariq M. Rana. Dynamics of the human and viral m^6A RNA methylomes during HIV-1 infection of T cells. *Nature Microbiology*, 1(4):16011, February 2016.

[294] Edward M. Kennedy, Hal P. Bogerd, Anand V.R. Kornepati, Dong Kang, Delta Ghoshal, Joy B. Marshall, Brigid C. Poling, Kevin Tsai, Nandan S. Gokhale, Stacy M. Horner, and Bryan R. Cullen. Posttranscriptional m^6A Editing of HIV-1 mRNAs Enhances Viral Gene Expression. *Cell Host & Microbe*, 19(5):675–685, May 2016.

[295] Nagaraja Tirumuru, Boxuan Simen Zhao, Wuxun Lu, Zhike Lu, Chuan He, and Li Wu. N^6-methyladenosine of HIV-1 RNA regulates viral infection and HIV-1 Gag protein expression. *eLife*, 5, 2016.

[296] Michael A. Weiss and Narendra Narayana. RNA recognition by arginine-rich peptide motifs. *Biopolymers*, 48(2-3):167–180, 1998.

[297] Rajan Lamichhane, John A. Hammond, Raymond F. Pauszek, Rae M. Anderson, Ingemar Pedron, Edwin van der Schans, James R. Williamson, and David P. Millar. A DEAD-box protein acts through RNA to promote HIV-1 Rev-RRE assembly. *Nucleic Acids Research*, 45(8):4632–4641, May 2017.

[298] Ibrahim I. Cisse, Ignacio Izeddin, Sebastien Z. Causse, Lydia Boudarene, Adrien Senecal, Leila Muresan, Claire Dugast-Darzacq, Bassam Hajj, Maxime Dahan, and Xavier Darzacq. Real-time dynamics of RNA polymerase II clustering in live human cells. *Science (New York, N.Y.)*, 341(6146):664–667, August 2013.

[299] Jesper V. Olsen, Blagoy Blagoev, Florian Gnad, Boris Macek, Chanchal Kumar, Peter Mortensen, and Matthias Mann. Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell*, 127(3):635–648, November 2006.

[300] Jesper V. Olsen, Michiel Vermeulen, Anna Santamaria, Chanchal Kumar, Martin L. Miller, Lars J. Jensen, Florian Gnad, Jürgen Cox, Thomas S. Jensen, Erich A. Nigg, Søren Brunak, and Matthias Mann. Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Science Signaling*, 3(104):ra3, January 2010.

[301] Noah Dephoure, Chunshui Zhou, Judit Villén, Sean A. Beausoleil, Corey E. Bakalarski, Stephen J. Elledge, and Steven P. Gygi. A quantitative atlas of mitotic phosphorylation. *Proceedings of the National Academy of Sciences of the United States of America*, 105(31):10762–10767, August 2008.

[302] Viveka Mayya, Deborah H. Lundgren, Sun-Il Hwang, Karim Rezaul, Linfeng Wu, Jimmy K. Eng, Vladimir Rodionov, and David K. Han. Quantitative phosphoproteomic analysis of T cell receptor signaling reveals system-wide modulation of protein-protein interactions. *Science Signaling*, 2(84):ra46, August 2009.

[303] Kristoffer T. G. Rigbolt, Tatyana A. Prokhorova, Vyacheslav Akimov, Jeanette Henningsen, Pia T. Johansen, Irina Kratchmarova, Moustapha Kassem, Matthias Mann, Jesper V. Olsen, and Blagoy Blagoev. System-wide temporal characterization of the proteome and phosphoproteome of human embryonic stem cell differentiation. *Science Signaling*, 4(164):rs3, March 2011.

[304] Shao-En Ong, Gerhard Mittler, and Matthias Mann. Identifying and quantifying in vivo methylation sites by heavy methyl SILAC. *Nature Methods*, 1(2):119–126, November 2004.

[305] Michel Soulard, Veronique Della Valle, Mikkiko C. Siomi, Serafin Pinol-Roma, Chantal Bauvy, Michel Bellini, Jean-Claude Lacroix, Guillaume Monod, Gidden Dreyfuss, Christian-Jacques Larsen, and others. hnRNP G: sequence and characterization of a glycosylated RNA-binding protein. *Nucleic acids research*, 21(18):4210–4217, 1993.

[306] Ivo A. Hendriks, Rochelle C. J. D'Souza, Bing Yang, Matty Verlaan-de Vries, Matthias Mann, and Alfred C. O. Vertegaal. Uncovering global SUMOylation signaling networks in a site-specific manner. *Nature Structural & Molecular Biology*, 21(10):927–936, October 2014.

[307] Francis Impens, Lilliana Radoshevich, Pascale Cossart, and David Ribet. Mapping of SUMO sites and analysis of SUMOylation changes induced by external stimuli.

Proceedings of the National Academy of Sciences of the United States of America, 111(34):12432–12437, August 2014.

[308] Zhenyu Xiao, Jer-Gung Chang, Ivo A. Hendriks, Jón Otti Sigurdsson, Jesper V. Olsen, and Alfred C. O. Vertegaal. System-wide Analysis of SUMOylation Dynamics in Response to Replication Stress Reveals Novel Small Ubiquitin-like Modified Target Proteins and Acceptor Lysines Relevant for Genome Stability. *Molecular & cellular proteomics: MCP*, 14(5):1419–1434, May 2015.

[309] Woong Kim, Eric J. Bennett, Edward L. Huttlin, Ailan Guo, Jing Li, Anthony Possemato, Mathew E. Sowa, Ramin Rad, John Rush, Michael J. Comb, J. Wade Harper, and Steven P. Gygi. Systematic and quantitative assessment of the ubiquitin-modified proteome. *Molecular Cell*, 44(2):325–340, October 2011.

[310] Lou K. Povlsen, Petra Beli, Sebastian A. Wagner, Sara L. Poulsen, Kathrine B. Sylvestersen, Jon W. Poulsen, Michael L. Nielsen, Simon Bekker-Jensen, Niels Mailand, and Chunaram Choudhary. Systems-wide analysis of ubiquitylation dynamics reveals a key role for PAF15 ubiquitylation in DNA-damage bypass. *Nature Cell Biology*, 14(10):1089–1098, October 2012.

[311] Y. R. Li, O. D. King, J. Shorter, and A. D. Gitler. Stress granules as crucibles of ALS pathogenesis. *The Journal of Cell Biology*, 201(3):361–372, April 2013.

[312] I. Collins, A. Weber, and D. Levens. Transcriptional consequences of topoisomerase inhibition. *Molecular and Cellular Biology*, 21(24):8437–8451, December 2001.

[313] C. Cassé, F. Giannoni, V. T. Nguyen, M. F. Dubois, and O. Bensaude. The transcriptional inhibitors, actinomycin D and alpha-amanitin, activate the HIV-1 promoter and favor phosphorylation of the RNA polymerase II C-terminal domain. *The Journal of Biological Chemistry*, 274(23):16097–16106, June 1999.

[314] Roland Schöller, Ignasi Forné, Tobias Straub, Amelie Schreieck, Yves Texier, Nilay Shah, Tim-Michael Decker, Patrick Cramer, Axel Imhof, and Dirk Eick. Heptad-Specific Phosphorylation of RNA Polymerase II CTD. *Molecular Cell*, 61(2):305–314, January 2016.

[315] Md Sohail Akhtar, Martin Heidemann, Joshua R. Tietjen, David W. Zhang, Rob D. Chapman, Dirk Eick, and Aseem Z. Ansari. TFIIH kinase places bivalent marks on the carboxy-terminal domain of RNA polymerase II. *Molecular Cell*, 34(3):387–393, May 2009.

[316] Kira Glover-Cutter, Stéphane Larochelle, Benjamin Erickson, Chao Zhang, Kevan Shokat, Robert P. Fisher, and David L. Bentley. TFIIH-associated Cdk7 kinase functions in phosphorylation of C-terminal domain Ser7 residues, promoter-proximal pausing, and termination by RNA polymerase II. *Molecular and Cellular Biology*, 29(20):5455–5464, October 2009.

[317] Nadine Czudnochowski, Christian A. Bösken, and Matthias Geyer. Serine-7 but not serine-5 phosphorylation primes RNA polymerase II CTD for P-TEFb recognition. *Nature Communications*, 3:842, May 2012.

[318] Olivier Bensaude. Inhibiting eukaryotic transcription: Which compound to choose? How to evaluate its activity? *Transcription*, 2(3):103–108, May 2011.

[319] Nova Fong, Hyunmin Kim, Yu Zhou, Xiong Ji, Jinsong Qiu, Tassa Saldi, Katrina Diener, Ken Jones, Xiang-Dong Fu, and David L. Bentley. Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. *Genes & Development*, 28(23):2663–2676, December 2014.

[320] Amy Pandya-Jones and Douglas L. Black. Co-transcriptional splicing of constitutive and alternative exons. *RNA (New York, N.Y.)*, 15(10):1896–1908, October 2009.

[321] Pierre C. Havugimana, G. Traver Hart, Tamás Nepusz, Haixuan Yang, Andrei L. Turinsky, Zhihua Li, Peggy I. Wang, Daniel R. Boutz, Vincent Fong, Sadhna Phanse, Mohan Babu, Stephanie A. Craig, Pingzhao Hu, Cuihong Wan, James Vlasblom, Vaqaar-un-Nisa Dar, Alexandr Bezginov, Gregory W. Clark, Gabriel C. Wu, Shoshana J. Wodak, Elisabeth R. M. Tillier, Alberto Paccanaro, Edward M. Marcotte, and Andrew Emili. A census of human soluble protein complexes. *Cell*, 150(5):1068–1081, August 2012.

[322] Edward L. Huttlin, Raphael J. Bruckner, Joao A. Paulo, Joe R. Cannon, Lily Ting, Kurt Baltier, Greg Colby, Fana Gebreab, Melanie P. Gygi, Hannah Parzen, John Szpyt, Stanley Tam, Gabriela Zarraga, Laura Pontano-Vaites, Sharan Swarup, Anne E. White, Devin K. Schweppe, Ramin Rad, Brian K. Erickson, Robert A. Obar, K. G. Guruharsha, Kejie Li, Spyros Artavanis-Tsakonas, Steven P. Gygi, and J. Wade Harper. Architecture of the human interactome defines protein communities and disease networks. *Nature*, 545(7655):505–509, 2017.

[323] Jacob C. Schwartz, Christopher C. Ebmeier, Elaine R. Podell, Joseph Heimiller, Dylan J. Taatjes, and Thomas R. Cech. FUS binds the CTD of RNA polymerase II and regulates its phosphorylation at Ser2. *Genes & Development*, 26(24):2690–2695, December 2012.