CHARACTERIZATION OF R-LOOP-INTERACTING PROTEINS IN EMBRYONIC STEM CELLS

A Dissertation Presented

By

TONG WU

Submitted to the Faculty of the

University of Massachusetts Graduate School of Biomedical Sciences, Worcester

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

October 30, 2021

Interdisciplinary Graduate Program

CHARACTERIZATION OF R-LOOP-INTERACTING PROTEINS IN EMBRYONIC STEM CELLS

A Dissertation Presented

By

TONG WU

This work was undertaken in the Graduate School of Biomedical Sciences

Interdisciplinary Graduate Program

Under the mentorship of

Thomas Fazzio, Ph.D., Thesis Advisor

Craig Peterson, Ph.D., Member of Committee

Erik Sontheimer, Ph.D., Member of Committee

Oliver Rando, M.D., Ph.D., Member of Committee

Amity Manning, Ph.D., External Member of Committee

Paul Kaufman, Ph.D., Chair of Committee

Mary Ellen Lane, Ph.D.,

Dean of the Graduate School of Biomedical Sciences

October 30, 2021

ACKNOWLEDGEMENTS

I would like to thank all the people who have provided suggestions, support, and help through graduate school. Most importantly, my mentor Tom Fazzio, who showed me how to do experiments, read journals, write research proposals, and give a talk, basically all the aspects of being a researcher. Also, I want to thank his support when it comes to life problems. I would like to thank the past and present members of the Fazzio lab, who have offered help with my experiments, research discussion, and scientific writing. All the past and present members of the Benanti lab for their suggestions and sharing of reagents. Thank you to my qualifying examination, TRAC, and dissertation examination committee members, Paul Kaufman, Craig Peterson, Erik Sontheimer, Oliver Rando, Sharon Cantor, Amity Manning. They have offered me invaluable scientific input and career guidance. I also thank members from MCCB and GSBS for scientific suggestions and career support.

It’s my great pleasure to collaborate with Feixia Chu and Jennifer Nance, their outstanding works are included in this thesis. I also feel great talking about science, refining presentations, and having fun together with my friends.

Finally, I would like to express my gratitude to my family, especially my partner Shanshan Guan. Even in different scientific fields, we discuss our science together. I am very grateful to have you to make every day the best.

ABSTRACT

RNAs associate with chromatin through various ways and carry out diverse functions. One mechanism by which RNAs interact with chromatin is by the complementarity of RNA with DNA, forming a three-stranded nucleic acid structure named R-loop. R-loops have been shown to regulate transcription initiation, RNA modification, and immunoglobulin class switching. However, R-loops accumulated in the genome can be a major source of genome instability, meaning that they must be tightly regulated. This thesis aims to identify R-loop-binding proteins systemically and study their regulation of R-loops.

Using immunoprecipitation of R-loops followed by mass spectrometry, with or without crosslinking, a total of 364 proteins were identified. Among them RNA-interacting proteins were prevalent, including some already known R-loop regulators. I found that a large fraction of the R-loop interactome consists of proteins localized to the nucleolus. By examining several DEAD-box helicases, I showed that they regulate rRNA processing and a shared set of mRNAs. Investigation of an R-loop-interacting protein named CEBPZ revealed its nucleolar localization, its depletion caused down-regulation of R-loops associated with rRNA and mRNA. Characterization of the genomic distribution of CEBPZ revealed its colocalization with insulator-regulator CTCF. When studying if CEBPZ recruits CTCF, I found that instead of regulating CTCF binding, CEBPZ depletion has a major effect on the performance of CUT&RUN, a technique for identifying DNA

binding sites of proteins. How CEBPZ affects CUT&RUN is still under investigation, the study of which may help us understand the roles of CEBPZ in regulation of global chromatin structure and genome integrity.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF COPYRIGHTED MATERIALS

The figures and content of the Chapter II has been published under the following citation:

Wu T., Nance J., Chu F., Fazzio T.G. Characterization of R-Loop-Interacting Proteins in Embryonic Stem Cells Reveals Roles in rRNA Processing and Gene Expression. *Mol Cell Proteomics*. 2021 Aug 31;20:100142. doi: 10.1016/j.mcpro.2021.100142.

LIST OF ABBREVIATIONS

3D genome	Three-dimensional genome
4SU	4-thiouridine
5’ ETS	5’ external transcribed spacer
AID	Auxin-inducible degron
ATAC-seq	Assay for transposase-accessible chromatin with high-throughput sequencing
ChIP-seq	Chromatin immunoprecipitation followed by sequencing
CID	Collision-induced dissociation
co-IP	Co-immunoprecipitation
CUT&RUN	Cleavage under targets and release using nuclease
DRIP-seq	DNA-RNA immunoprecipitation sequencing
DSBs	Double-strand breaks
esiRNA	Endoribonuclease-prepared small interfering RNA
G4s	G-quadruplexes
GO term	Gene ontology term
hnRNP	Heterogeneous nuclear ribonucleoprotein
IAA	Indole-3-acetic acid
IF	Immunofluorescence staining
LC-MS/MS	Liquid chromatography and tandem mass spectrometry
lncRNA	Long non-coding RNA
m6A	N6-methyladenosine

mESCs	Mouse embryonic stem cells
RBP	RNA binding protein
RDHs	RNA-DNA hybrids
RNAP	RNA polymerase
RNAP I	RNA polymerase I
RNAP II	RNA polymerase II
RNAP III	RNA polymerase III
rRNA	Ribosomal RNA
SILAC	Stable isotope labeling by amino acids in cell culture
ssDNA	Single-stranded DNA
ssRNA	Single-stranded RNA
TAD	Topologically associated domain
TRCs	Transcription-replication conflicts
XCI	X-chromosome inactivation
Xi	Inactive X chromosome
Xist	X-inactivation specific transcript

CHAPTER I: Introduction

Characterization of chromatin-associated RNAs

RNAs have long been known to be co-purified with chromatin (Bonner *et al.*, 1968; Kanehisa *et al.*, 1971; Mayfield and Bonner, 1971). After mechanical homogenization, salt wash, and centrifugation, RNAs can still be found within the chromatin fraction, indicating that some transcripts stably interact with chromatin instead of floating in nucleoplasm. These chromatin-associated RNAs vary a lot, depending on the tissues from which they were purified from. They tend to hybridize with the chromatin and DNA isolated from the same tissues rather than different tissues (Kanehisa *et al.*, 1971; Mayfield and Bonner, 1971), suggesting that these chromatin-associated RNAs are products of tissue-specific transcription.

It is reasonable to assume that the chromatin-associated RNAs are nascent transcripts undergoing transcription on their templates, still associated with RNA polymerase (RNAP). However, the development of new techniques to characterize RNAs has depicted a more complicated picture, showing RNAs associated with DNA through direct RNA-DNA interaction and RNAs associated with chromatin through indirect interactions via various RNA binding proteins (RBPs).

One of the most well-known examples of direct interactions of RNA with chromatin is RNA-DNA hybrids (RDHs). Once the single-stranded RNA (ssRNA) anneals with its complementary single-stranded DNA (ssDNA), it forms RDHs that

are more thermodynamically stable than dsDNA (Lesnik and Freier, 1995) and forms structures that are in between B-form of dsDNA and A-form of dsRNA. Along with the ssDNA that is displaced by the formation of RDHs, the whole structure is called an R-loop, named after a similar nucleic acid structure named D-loop, which forms when ssDNA invades dsDNA (Thomas, White and Davis, 1976). Since the discovery of R-loops, they have been considered as by-products of transcription and one of the major causes of genome instability (Aguilera and García-Muse, 2012). However, R-loops have also been known to regulate processes such as immunoglobulin class switching (Yu *et al.*, 2003), DNA replication (Lombraña *et al.*, 2015), and transcription initiation (Ginno *et al.*, 2012). In recent years, new roles of R-loop have been found, including regulation of DNA repair (Marnef and Legube, 2021) and RNA modification (Marnef and Legube, 2020).

Another form of direct interaction of RNA with chromatin occurs via formation of triplex nucleic structures, specifically, RNA-dsDNA triplexes. These structures form based on the Hoogsteen hydrogen bonds between the ssRNA and the purine-rich strand of dsDNA (Hoogsteen, 1963; Morgan and Wells, 1968; Robles *et al.*, 2005; Bacolla, Wang and Vasquez, 2015), rather than the Watson-Crick hydrogen bonds that are the driving force of RNA-DNA interaction in R-loops. This results in ssRNA sitting in the dsDNA major groove and interacting with one strand of DNA (Goñi, de la Cruz and Orozco, 2004). Like R-loops, RNA-dsDNA triplexes have been found to regulate processes including transcription initiation

(Zhou, Giles and Felsenfeld, 2019), DNA methylation (Schmitz *et al.*, 2010), and recruitment of chromatin remodeling complexes (Grote and Herrmann, 2013).

Some RNAs associate with chromatin in a protein-dependent manner, making them able to survive harsh chromatin extraction. One of the most commonly found chromatin-associated RNAs are pre-mRNAs that are tethered to chromatin by RNA polymerase II (RNAP II). Once they exit RNAP II, pre-mRNAs are coated by factors that lead to their maturation, including factors that carry out capping, splicing, and polyadenylation (Bentley, 2014). Coupling ongoing transcription with splicing has been shown to be essential to achieve alternative splicing (Naftelberg *et al.*, 2015). Regulation of transcription by spliceosome factors that interact with transcription elongation factors has also been observed (Fong and Zhou, 2001).

RNAs can associate with chromatin in an RBP-dependent manner without directly interacting with genomic DNA. One of the most commonly known examples occurs during X-chromosome inactivation (XCI). In female mammals, one of the two X chromosomes needs to be silenced to achieve dosage compensation of X chromosomal genes between females and males (Lyon, 1961). This process is achieved by binding of the inactive X chromosome (Xi) by the noncoding X-inactivation specific transcript (Xist), which is transcribed exclusively from the Xi. Coating of the Xi with Xist RNA leads to depletion of the transcription machinery, loss of active histone markers, and gain of repressive histone markers on the Xi, finally turning it into a stably inactivated chromosome in female somatic

cells (Wutz, 2011). The mechanisms by which Xist interacts with Xi stably are still undergoing investigation, but several RBPs have been shown to bridge Xist RNA with Xi DNA. Heterogeneous nuclear ribonucleoprotein U (HNRNPU, also named SAF-A) is a protein with both DNA- and RNA- interacting activities. Several groups have shown that the RNA-binding domain of HNRNPU interacts with Xist while the DNA-binding domain interacts with AT-rich chromosomal regions on the Xi (Helbig and Fackelmayer, 2003; Hasegawa *et al.*, 2010). Lacking either of the domains or HNRNPU itself leads to a failure of Xist to localize to the X chromosome, indicating that HNRNPU is essential for Xist binding to Xi (Hasegawa *et al.*, 2010). YY1, a protein that also harbors both DNA- and RNA- binding domains, has been shown to be essential for bringing Xist to Xi (Jeon and Lee, 2011; Makhlouf *et al.*, 2014). Despite these RBP-mediated indirect interactions of RNA with chromatin, direct interaction of Xist RNA with Xi DNA by the formation of R-loops or triplex may also exist.

Formation and distribution of R-loops

R-loops are a subset of chromatin-associated RNA structures that form when ssRNA binds dsDNA in a sequence-specific manner with one strand of DNA being looped out. R-loops are often generated during transcription when nascent RNAs transcribed by RNAPs thread back and anneal with the DNA templates from

which they are transcribed. Transcription-coupled RDHs are the major source of R-loops found in cells (Belotserkovskii *et al.*, 2018).

The formation and stability of R-loops are regulated by several nucleic acid features, including purine-rich sequences in nascent transcripts (Roy and Lieber, 2009), the formation of G-quadruplexes (G4s) on the free ssDNA (Duquette *et al.*, 2004), negative supercoiling (Stolz *et al.*, 2019), DNA nicks on the non-template strand (Allison and Wang, 2019), pausing of RNA polymerase, and lack of factors that inhibit or resolve R-loops (Figure 1.1). For example, RDHs composed of G-rich RNA/C-rich DNA exhibit higher stability than equivalent dsDNA composed of G-rich DNA/C-rich DNA, stabilizing the RDHs once they form (Gyi *et al.*, 1996). Moreover, the displaced G-rich ssDNA may form G4s, which will further stabilize the R-loop structures. Negative supercoiling, which contains a high-stress, high-energy state of the dsDNA, also favors R-loop formation as RDHs and the ssDNA of R-loops will adapt into a more relaxed state compared with negative supercoils (Stolz *et al.*, 2019).

In the past decade, genome-wide approaches have been developed to understand the distribution and dynamics of R-loops. These studies indicate that R-loops are highly abundant throughout the genome, especially on coding genes, in various organisms (Chen *et al.*, 2015; Sanz *et al.*, 2016; Wahba *et al.*, 2016; Lang *et al.*, 2017; Xu *et al.*, 2017). The first described genome-wide approach to map RDHs, known as DNA-RNA immunoprecipitation sequencing (DRIP-seq), used the RDH-specific monoclonal antibody S9.6 to detect R-loops, showing R-

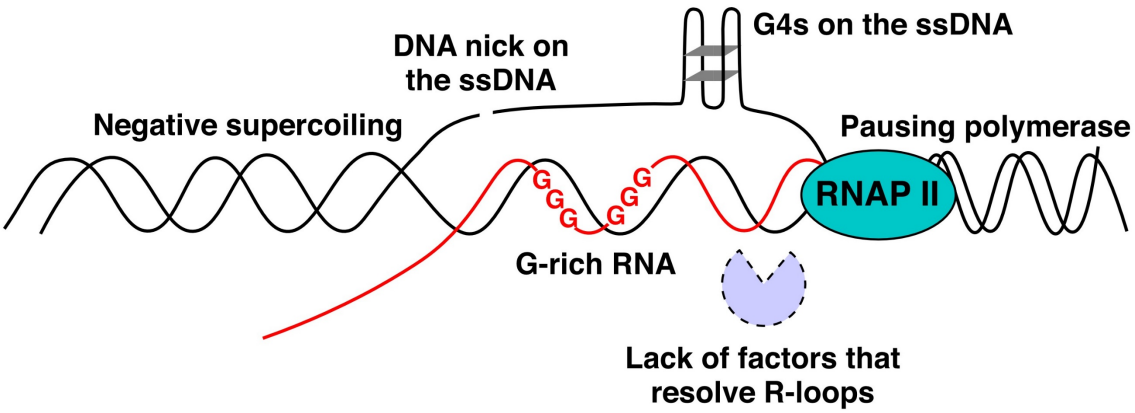

Figure 1.1. Nucleic acid features that favor R-loop formation.

G4s: G-quadruplexes. RNAP II: RNA polymerase II.

loop accumulation on unmethylated CpG island promoters (Boguslawski *et al.*, 1986; Ginno *et al.*, 2012). Various other R-loop-mapping methods have subsequently been developed. In addition to studies that use S9.6 to detect R-loops (Ginno *et al.*, 2013; El Hage *et al.*, 2014), some groups have focused on catalytically dead RNase H1, which binds specifically to RDHs but does not digest them (Chen *et al.*, 2017, 2019; Yan *et al.*, 2019). Although these studies vary somewhat, especially when comparing S9.6-based with dead RNase H1-based techniques (Chédin *et al.*, 2021), they all revealed that a large portion of R-loops co-localize with nascent transcripts, with high enrichment at promoter-proximal regions, lower (but still detectable) levels at transcription termination sites and

gene bodies, and little or no signal at intergenic regions (Sanz *et al.*, 2016; Chen *et al.*, 2017; Chédin *et al.*, 2021). In agreement with these findings, R-loops that form at transcription start and termination sites can regulate certain processes，including transcription initiation and termination. In particular, R-loops that form around promoters have been shown to promote transcription by inhibiting the binding of DNA methyltransferase enzymes, while R-loops at transcriptional pause sites recruit the helicase Senataxin (SETX) and exonuclease XRN2, inducing RNAP II release and transcription termination (Skourti-Stathaki, Proudfoot and Gromak, 2011; Ginno *et al.*, 2012).

Interactome and regulation of R-loops

Since the first description of R-loops in 1976 (Thomas, White and Davis, 1976), various proteins have been shown to bind to RDHs/R-loops and regulate their abundance. These include the nucleases RNase H1 (Stein and Hausen, 1969; Cerritelli and Crouch, 2009) and RNase H2 (Cornelio *et al.*, 2017) that specifically digest the RNA component of RDHs, topoisomerases Top1 and Top2 that inhibit R-loop formation by resolving negative supercoiling (El Hage *et al.*, 2010), helicases like Sen1 in yeast (Kim, Choe and Seo, 1999) and its homologue SETX in humans (Skourti-Stathaki, Proudfoot and Gromak, 2011) that resolve RDHs, and messenger ribonucleoprotein THO/TREX (Domínguez-Sánchez *et al.*, 2011) and splicing factor SRSF1 (Li and Manley, 2005), which bind to nascent

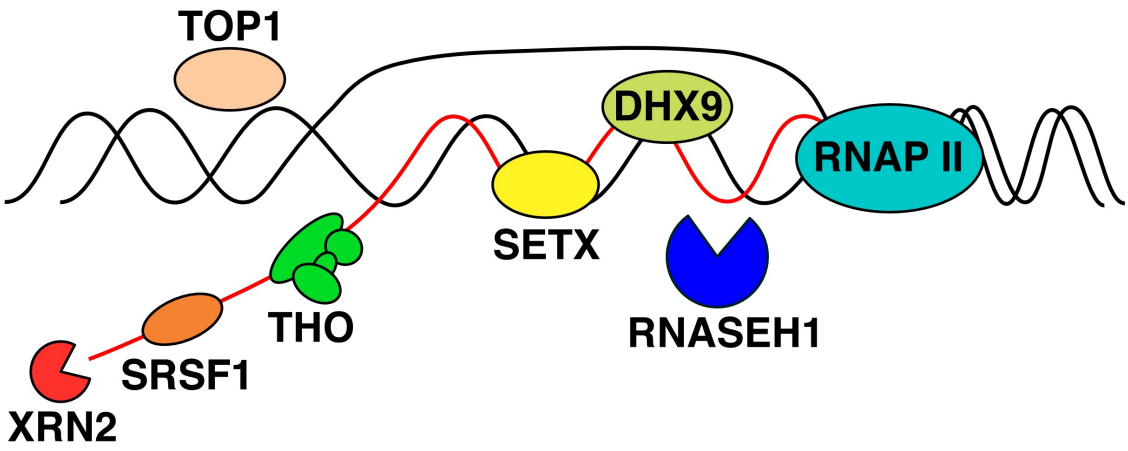

Figure 1.2. Factors that resolve R-loops or inhibit R-loop formation.

Nucleases that digest the RNA component of R-loops: RNASEH1, XRN2. Helicases that resolve RNA-DNA hybrids: SETX, DHX9. Factors that bind to mRNA to inhibit R-loop formation: THO, SRSF1. Topoisomerase that resolves negative supercoiling to inhibit R-loop formation: TOP1.

mRNA to inhibit R-loop formation (Figure 1.2). Moreover, R-loops have been shown to regulate the distribution of chromatin proteins and protein complexes in the genome, including chromatin remodeling complexes Tip60-p400 and PRC2 (Chen *et al.*, 2015) and DNA methyltransferase DNMT3B1 (Ginno *et al.*, 2012).

To identify R-loop-interacting proteins in a systematic way, the Gromak group used the S9.6 antibody to capture R-loops together with their binding proteins, followed by identification of these proteins by mass spectrometry (MS)

(Cristini *et al.*, 2018). A total of 469 proteins were identified, including heterogeneous nuclear ribonucleoproteins (hnRNPs) that function in RNA metabolism, DNA binding proteins like topoisomerase TOP1, splicing factors, helicases, and ribosomal RNA (rRNA) processing factors. The authors went on to characterize one of these interacting proteins, a helicase named DHX9. The study showed that DHX9 is a factor that inhibits R-loops formation, at least at certain genomic loci, in response to TOP1 inhibition. The R-loop regulatory activity of DHX9 *in vivo* agrees with previous research showing unwinding of R-loops by DHX9 *in vitro* (Chakraborty and Grosse, 2011), and further confirmed by research published later (Chakraborty, Huang and Hiom, 2018).

In a second study, a different assay was carried out to capture and identify RDH-interacting proteins from protein fractions (Wang *et al.*, 2018). In this assay, RDHs generated by annealing oligonucleotides *in vitro* were incubated with human B-cell extract to capture RDH-interacting proteins, followed by liquid chromatography and tandem mass spectrometry (LC-MS/MS). This study identified 803 proteins shared between two RDH baits, including known R-loop-regulators like helicase DDX5, splicing factor SFPQ, exonuclease XRN2, and the potential R-loop-regulator FUS, a protein that has both DNA- and RNA- binding activities (Yamaguchi and Takanashi, 2016).

By overlapping proteins identified in the two studies above, a total of 197 proteins were identified, indicating the reliability of both methods. However, a large group of proteins does not overlap between these two sets. This may be explained

by the fact that the *in vivo* assay captures not only proteins that bind directly to R-loops but also the ones within the sheared chromatin region, while the *in vitro* assay is unable to identify the R-loop-interacting proteins that bind to ssDNA or the entire structure of R-loops, as well as proteins that participate in R-loop regulation co-transcriptionally. Among the proteins that overlap between the two studies are various DEAD-box proteins with known or putative ATP-dependent RNA helicase activity (Linder *et al.*, 1989; Cordin *et al.*, 2006; Linder, 2006). DEAD-box proteins have been found to regulate mRNA transcription, RNA degradation, splicing, rRNA processing, and ribosome biogenesis (Linder, 2006). Several DEAD-box proteins have been confirmed to regulate R-loops independently of these proteomic studies examining the R-loop interactome (Hodroj *et al.*, 2017; Song *et al.*, 2017; Ribeiro de Almeida *et al.*, 2018).

Regulation of transcription by R-loops

RNA polymerase I (RNAP I) transcribes 18S and 28S rRNA in the nucleolus, which produces about 80% of all RNA in the cell (Harvey *et al.*, 2000). Nascent rRNA anneals with DNA regions it is transcribed from to form R-loop structures, which may contribute to the high abundance of R-loops observed in the nucleolus by S9.6 immunofluorescence staining (IF) (El Hage *et al.*, 2010; García-Rubio *et al.*, 2015; Shen *et al.*, 2017). Depletion of topoisomerase Top1 in yeast causes accumulation of R-loops at the 5’ end of ribosomal DNA (rDNA) loci, which induces

pileup of RNAP I and impairs rRNA synthesis (El Hage *et al.*, 2010). Similar observation were made with RNA polymerase III (RNAP III) transcribed transfer RNAs (tRNAs) upon loss of Top1 (El Hage *et al.*, 2014).

At RNAP II-transcribed genes, on which most genome-wide studies focus, Top1 inhibition was observed to pause RNAP II at actively transcribed genes (Baranello *et al.*, 2009). However, mechanisms by which R-loops regulate mRNA transcription are more complicated, involving inhibition and activation of transcription. At promoters that contain CpG islands, G-rich transcripts tend to form R-loops, protecting these sites from binding of DNA methyltransferase enzymes, thus preventing gene silencing (Ginno *et al.*, 2012). It has also been shown that R-loop formation can be induced by antisense RNAs. As one example, antisense RNA-mediated R-loop formation has been observed at the promoter of the tumor suppressor gene *TCF21*. Furthermore, stress response protein GADD45A has been shown to bind R-loops and recruit the TET1 enzyme, which demethylates DNA to control the expression of *TCF21* in a cell cycle-dependent manner (Arab *et al.*, 2019). In mouse embryonic stem cells (mESCs), R-loops at promoter-proximal regions recruit the activating chromatin complex Tip60-p400 and inhibit the binding of polycomb repressive complex 2 (PRC2), which may contribute to the differentiation defects observed in mESCs with disrupted R-loops (Chen *et al.*, 2015). However, R-loops may have opposite functions regarding the specific locus and cell line they were examined. For example, the long non-coding RNA (lncRNA) *ANRASSF1*, the antisense RNA of the *RASSF1A* gene, can induce R-loop

formation and recruit PRC2 complexes to the *RASSF1A* promoter to suppress gene expression in HeLa cells (Beckedorff *et al.*, 2013).

Regulation of transcription termination by R-loops has been observed at different genomic loci in different cells types. In HeLa cells, G-rich pause sequences downstream of polyadenylation sites at the *ACTB* and *MAZ4* genes favor the formation of R-loops, which further induces RNAP II pausing (Yanling Zhao *et al.*, 2016). Paused RNAP II recruits SETX, which resolves RDHs to relieve ssRNA that is digested by exoribonuclease XRN2. This results in RNAP II release and efficient termination (Skourti-Stathaki, Proudfoot and Gromak, 2011). Moreover, R-loops can facilitate transcription termination by epigenetic regulation. For example, antisense transcription can be induced by stabilized R-loops at termination regions of *ACTB* gene, leading to the formation of dsRNA that recruits histone methyltransferase G9a. The repressive histone marker H3K9me2, deposited by G9a, then recruits heterochromatin protein 1γ (HP1γ), facilitating the formation of heterochromatin and pausing of RNAP II (Skourti-Stathaki, Kamieniarz-Gdula and Proudfoot, 2014). In recent years, evidence has emerged linking R-loops with N6-methyladenosine (m6A) modification, the most abundant RNA modification on mRNA. Positive correlation of m6A-containing transcripts with their tendency to form R-loops has been shown in human induced pluripotent stem cells (Abakir *et al.*, 2019). Moreover, upon depletion of METTL3, the protein that converts adenosine to m6A, reduction of m6A and R-loop levels was observed

at termination regions where they co-exist, causing RNAP II readthrough and inefficient transcription termination (Niehrs and Luke, 2020).

Roles of R-loops in genome stability

Independent of the regulatory roles of R-loops in transcription, R-loops have long been considered a major source of genome instability. For example, mutations in the ribonucleoprotein complex THO/TREX induce higher levels of R-loops, leading to hyperrecombination and chromosome loss (Huertas and Aguilera, 2003). In HeLa cells, hyperactivation of NF-κB has been shown to induce R-loop accumulation and formation of double-strand breaks (DSBs), which can be rescued by overexpressing the RDH-specific nuclease RNase H1 (He *et al.*, 2021).

The mechanisms by which R-loops lead to genome instability have been thoroughly dissected over the past 20 years. The ssDNA part of R-loops can be the cause of R-loop-induced genome instability, as ssDNA serves as a substrate for mutagen activation-induced cytidine deaminase (AID) (Gómez-González and Aguilera, 2007). AID deaminates cytosines into uracil, and may further induce DNA breaks and chromosome translocations, as the mis-incorporated uracil may be cleaved to leave a break (Robbiani *et al.*, 2009). A more detrimental effect of R-loops may come from their ability to block the DNA replication machinery, due to the pausing RNAP II that is associated with R-loops. Transcription-replication conflicts (TRCs) may lead to the collapse of replication forks and chromosome

rearrangements (Hamperl and Cimprich, 2016). In one study performed in HEK293 cells, an episomal system was introduced to enable the dissection of TRCs. This artificial system shows that in head-on collisions of the RNAP II with replication forks, the ATR-dependent DNA damage response system is triggered and leads to increased R-loops (Hamperl *et al.*, 2017). Several groups have shown TRCs to be sources of R-loop-induced DNA damage (Prado and Aguilera, 2005; Gan *et al.*, 2011; Helmrich, Ballarino and Tora, 2011). However, there is still no clear understanding of the frequency of TRCs occurrence in the genome, nor have mechanisms been described that explain how they would induce DSBs in an R-loop-dependent way.

Although R-loops have long been considered a source of genome instability, recently, several studies indicate that R-loops can form at the sites of DSBs in a transcription-dependent manner to regulate DNA repair. In yeast, RNAP II is recruited to the 3' ssDNA overhangs of DSBs after end resection is complete. The nascent RNA transcribed from RNAP II can then hybridize with the ssDNA to form RDHs, which must then be degraded by RNase H1 and RNase H2 for replication protein A (RPA) to be loaded for additional steps of the DNA repair process (Ohle *et al.*, 2016). In a human epithelial cell line, stalling RNAP II induces R-loops formation around DSB sites, which have been shown to recruit RAD52 and facilitate endonuclease XPG-dependent removal of R-loops. This process is required for the initiation of transcription-dependent homologous recombination repair. It is also worth noting that this repair pathway may occur in around 5% of

total DSBs, depending on the level of transcription around these sites once DSBs form (Yasuhara *et al.*, 2018).

Functions of R-loops in mouse embryonic stem cells

Several groups have shown that disruption of R-loops in mESCs affects recruitment of chromatin remodeling complexes like PRC1, PRC2, and Tip60-p400, which in turn affects expression of genes bound by these complexes (Chen *et al.*, 2015; Skourti-Stathaki *et al.*, 2019). Because of the regulation of mESC differentiation by these remodeling complexes (Fazzio, Huff and Panning, 2008; Walker *et al.*, 2010; Chen *et al.*, 2013), disruption of R-loops also causes differentiation defects (Chen *et al.*, 2015). Moreover, mutation of death inducer obliterator 3 (DIDO3), a protein that interacts with helicase DHX9 and regulates R-loops, causes differentiation defects in mESCs (Fütterer *et al.*, 2021). All these studies suggest that R-loop-regulation plays an essential role in the homeostasis of mESCs, and that disruption of R-loops will cause observable biological consequences, making mESCs a good platform for studying R-loop functions.

mESC lines were first generated from the inner cell mass of mouse blastocyst stage embryos in 1981 by two different groups. These cell lines were maintained in defined medium supplemented with serum (Evans and Kaufman, 1981), or in a conditioned medium (Martin, 1981). Both studies required feeder cells for the maintenance of the ESCs. These ESCs grow in colonies with rapid

and stable replication, allowing them to be maintained *in vitro* indefinitely without directed or spontaneous mutations that are required in most other primary cell types. Once these cell lines are passed without feeder cells, they spontaneously differentiate into three-dimensional structures named embryoid bodies that consist of three embryonic germ layers, indicating their differentiation ability (Koike *et al.*, 2007). Moreover, once injected into mice, these cell lines differentiate into teratomas, germ cell tumors made up of mature cell types from three primary germ layers that are differentiated from the ESCs (Peterson *et al.*, 2012).

As characterized by these early research and others, mESCs are defined by two features: self-renewal and pluripotency. Self-renewal indicates the ability of ESCs to proliferate indefinitely under appropriate culture conditions without differentiation, while pluripotency is the ability of ESCs to differentiate into three germ layers, endoderm, mesoderm, and ectoderm, which contribute to both somatic and germinal lineages. It is worth noting that mESCs are pluripotent but not totipotent, as mESCs do not contribute to extraembryonic lineages (Beddington and Robertson, 1989; Condic, 2014). Though mESCs can be maintained *in vitro* in a state of pluripotency with embryonic fibroblasts as feeder cells (Evans and Kaufman, 1981; Martin, 1981; Suda *et al.*, 1987), it was later found that mESCs can be cultured on gelatin-coated plates, in serum-containing medium supplemented by leukemia inhibitory factor (LIF), a cytokine secreted by feeder cells that inhibit differentiation (Koopman and Cotton, 1984; Williams *et al.*, 1988). LIF is a member of the interleukine-6 cytokine family (IL-6). LIF binds to its receptor,

which activates Janus kinase (JAK) that phosphorylates signal transducer and activator of transcription 3 (STAT3). STAT3 dimerizes and enters the nucleus to regulate gene regulatory networks through several pathways. They include activation of MYC, and activation of SOX2 and POU5F1 (Oct-4) through KLF4, which are essential for ESC self-renewal and pluripotency (Cartwright *et al.*, 2005; Hall *et al.*, 2009; Niwa *et al.*, 2009; Tang and Tian, 2013).

Transcription factor CEBPZ as a possible R-loop regulator

CCAAT Enhancer Binding Protein Zeta (CEBPZ) is a transcription factor that specifically recognizes the CCAAT sequence of promoters. CEBPZ was linked to R-loops by research from Barbieri *et al.*, which showed that CEBPZ and METTL3 co-localize extensively throughout the genome, as measured by chromatin immunoprecipitation followed by sequencing (ChIP-seq). CEBPZ mediates the recruitment of METTL3 to its promoters. Depletion of CEBPZ leads to reduction of m6A modification on transcripts associated with METTL3-associated promoters, as METTL3 is known to carry out m6A modification (Barbieri *et al.*, 2017; Zaccara, Ries and Jaffrey, 2019). In recent years, several groups have shown that m6A-containing transcripts promote R-loop formation (Abakir *et al.*, 2019; Yang *et al.*, 2019; Marnef and Legube, 2020), suggesting the regulation of R-loops by m6A incorporation and METTL3 binding. These studies indicate a

possible connection between CEBPZ and R-loops through the regulation of m6A abundance.

CEBPZ was first characterized in 1990, in which it was identified as a 114-kD transcription factor that binds to the promoter of the human *hsp70* gene. It binds specifically to the CCAAT sequence of the promoter and drives *hsp70* transcription in a CCAAT sequence-dependent manner, which gives its name CCAAT binding factor (CBF) (Lum *et al.*, 1990). The N-terminus of CEBPZ interacts directly with adenovirus E1a protein and tumor suppressor p53 (Lum *et al.*, 1992; Agoff *et al.*, 1993), these interactions have been later confirmed to drive the E1a-induced *hsp70* promoter activation and p53-mediated *hsp70* repression (Agoff and Wu, 1994; Chae, Yun and Shin, 2005). It was later found that CEBPZ recruitment to the *hsp70* promoter depends on its interaction with the nuclear factor Y (NF-Y), a trimeric complex that is also known to bind to the CCAAT sequence (Imbriano *et al.*, 2001). In 1996, a mouse ortholog of CEBPZ was identified and shown to localize to the nucleus. It has > 80% amino acid sequence similarity with the human CEBPZ (Hoeppner *et al.*, 1996). MAK21P, a budding yeast homolog of human and mouse CEBPZ, has been shown to be essential for both 60S ribosomal subunit biogenesis and cell growth (Edskes, Ohtake and Wickner, 1998).

Though it has the name CCAAT Enhancer Binding Protein Zeta, CEBPZ should not be regarded as belonging to the C/EBP family, as it lacks the basic leucine zipper domain that is shared by all members of that family. Phylogenetic analysis based on protein sequences of the C/EBP family and CEBPZ shows that

CEBPZ forms its own clade, indicating a different origin than CEBPA, CEBPB, CEBPD, CEBPE, CEBPG, and DDIT3 (also known as CHOP and occasionally also called CEBPZ) that belong to the C/EBP family (Pulido-Salgado, Vidal-Taboada and Saura, 2015).

Insulator-binding protein CTCF and its regulation by R-loops

CCCTC-binding factor (CTCF) is a zinc finger protein that is highly conserved in higher eukaryotes (Heger *et al.*, 2012). CTCF contains eleven zinc figures that are essential for its binding to the genome (Vostrov, Taheny and Quitschke, 2002), which are highly conserved between mouse and human (Ohlsson, Renkawitz and Lobanenkov, 2001). CTCF was first identified as a transcriptional repressor of the *c-myc* gene in chicken (Klenova *et al.*, 1993). Now CTCF has been assigned multiple functions, including transcription activation/repression, insulation, and regulation of the three-dimensional genome (3D genome).

The insulation function of CTCF was first characterized by the Felsenfeld group. They described a 42-bp sequence of chicken β-globin locus that binds CTCF. Upon CTCF binding, it blocks the interaction of the enhancer and promoter, thus inhibiting transcription (Bell, West and Felsenfeld, 1999). Later, Felsenfeld and several other groups showed that mouse insulin-like growth factor 2 (*Igf2*) is controlled by the insulator activity of CTCF (Bell and Felsenfeld, 2000; Hark *et al.*,

2000; Szabó *et al.*, 2000). *Igf2* and *H19* share the same enhancer at the 3' end of *H19* (Leighton *et al.*, 1995). Upon methylation of regions between *Igf2* and *H19,* which are CTCF-binding sites, CTCF binding decreases while the interaction of enhancer and *Igf2* promoter increases, thus reversing the transcription inhibition caused by CTCF insulation.

The promoters of metazoans can be regulated by multiple clusters of enhancers located at long distances relative to the promoters (Levine, Cattoglio and Tjian, 2014), some as far as several megabases (Amano *et al.*, 2009; Shi *et al.*, 2013). This suggests the formation of genome structures at higher orders to couple promoters with their enhancers despite their long-distance on the linearized genome. Indeed, the insulation activity of CTCF can be carried out through regulation of genome architecture. This is demonstrated by the finding that CTCF tethers the chicken β-globin insulator to subnuclear regions, creating independently looped domains such that an enhancer in one loop will not be able to interact with a promoter in another loop (Yusufzai *et al.*, 2004). CTCF-mediated long-range looping around the β-globin locus in mice can be disabled by depletion of CTCF or disruption of the CTCF binding site, causing changes in local histone modifications (Splinter *et al.*, 2006). More direct evidence that CTCF controls transcription by the formation of loops was shown by introducing an ectopic insulator in mice. Together with the endogenous insulator, the ectopic insulator was bound by CTCF, causing the two loci to come together to form a loop that sequestered intervening enhancers and leading to transcription repression (Hou

et al., 2008). On a genome-wide scale, chromatin conformation capture techniques have revealed 3D chromatin structures, including compartments, topologically associated domains (TADs), and loops (Lieberman-Aiden *et al.*, 2009; Dixon *et al.*, 2012; Rao *et al.*, 2014). Chromosome compartments are active (compartment A) and repressive (compartment B) regions on chromosomes that are megabases in scale. They represent chromosome regions that are functionally distinct and spatially separated (Lieberman-Aiden *et al.*, 2009). TADs are chromosome blocks that expand hundreds of kilobases in which sequences in the same TAD interact with each other at a higher frequency than with sequences outside of the TAD. The boundaries of TADs are usually occupied by CTCF and cohesin complexes (Dixon *et al.*, 2012; Rao *et al.*, 2014). CTCF and cohesin mediate formation of TADs and loops through a loop extrusion mechanism, in which cohesin rings slide through the genome to generate loop structures until blocked by CTCF on its binding sites (Fudenberg *et al.*, 2016). Disruption of TADs, either by depletion of CTCF or deletion of TAD boundaries, will cause misregulation of genes within and nearby TADs (Lupiáñez *et al.*, 2015; Nora *et al.*, 2017).

Recently, Luo *et al.* showed that CTCF binding at a subset of loci is mediated by R-loops, which is essential for CTCF binding and TAD formation. A HOXA locus-associated long noncoding RNA (lncRNA), *HOTTIP*, has been shown to induce R-loop formation at genomic sites where CTCF binds, for example, at TAD boundaries of β-catenin. Disruption of the R-loops at these sites by dCas9-guided RNase H reduced binding of *HOTTIP* and CTCF to the chromatin.

Moreover, TAD topology was impaired upon disruption of R-loops at its boundaries, with the expression of β-catenin impaired. This study connects the regulation of 3D genome by CTCF with the presence of R-loops (Luo *et al.*, 2020).

Summary and perspectives

Since the first description of R-loops (Thomas, White and Davis, 1976), more and more factors have been shown to regulate R-loop abundance, with the majority of them being factors that resolve R-loops or prevent their formation. There has long been a lack of systemic identification of R-loop-interacting proteins, until Cristini *et al.* reported a method using the S9.6 antibody to capture R-loops from fragmented chromatin followed by MS to identify its interacting proteins. However, R-loop-proximal chromatin with its binding proteins is very likely to be co-purified with R-loops, making it difficult to distinguish proteins that bind to R-loops from the ones that bind to close regions. Therefore, a more specific strategy needs to be carried out to study R-loop-interacting proteins, which is described in Chapter II. Using mESCs as the model cell line, I was able to identify a set of R-loop-interacting proteins using a stringent washing protocol. 4-thiouridine (4SU) labeling followed by ultraviolet B (312 nm) crosslinking enabled the identification of transiently or weakly R-loop-interacting proteins. I found that a large portion of the proteins identified localize to the nucleolus partially or completely. In Chapter III, I examined CEBPZ, which has been identified as an R-loop-interacting protein.

I found that it regulates rRNA- and mRNA-associated R-loops. Surprisingly, CEBPZ was shown to colocalize with CTCF at specific genomic regions, with their physical interaction characterized by immunoprecipitation. In Chapter IV, I will emphasize the findings in this thesis, discuss how these results help us understand the interactome and regulation of R-loops, and propose future directions of this research.

CHAPTER II: Characterization of R-loop-interacting DEAD-box proteins reveals roles in rRNA processing and gene expression

Preface

Data presented in this chapter are published in Molecular & Cellular Proteomics, with the title “Characterization of R-Loop-Interacting Proteins in Embryonic Stem Cells Reveals Roles in rRNA Processing and Gene Expression”, under the following citation:

Wu T., Nance J., Chu F., Fazzio T.G. Characterization of R-Loop-Interacting Proteins in Embryonic Stem Cells Reveals Roles in rRNA Processing and Gene Expression. *Mol Cell Proteomics*. 2021 Aug 31;20:100142. doi: 10.1016/j.mcpro.2021.100142.

Author contributions. Tong Wu, Jennifer Nance, Feixia Chu, and Thomas Fazzio designed experiments. Tong Wu performed most of the experiments. Jennifer Nance performed in-gel digestion of the co-immunoprecipitation samples, ran LC-MS/MS. Jennifer Nance and Feixia Chu performed peptide assignment.

Abstract

To identify factors that may bind and regulate R-loop accumulation or mediate R-loop-dependent functions, we used antibody to capture R-loops with their binding proteins followed by mass spectrometry, with and without RNA-protein crosslinking, to identify a stringent set of R-loop-binding proteins in mouse embryonic stem cells. We identified 364 R-loop-interacting proteins, which were highly enriched for proteins with predicted RNA-binding functions. A high portion of the identified proteins are nucleolar proteins, agree with the fact that ribosomal RNA transcription contributes to R-loop formation. We characterized several R-loop-interacting proteins of the DEAD-box family of RNA helicases and found that these proteins localize to the nucleolus and, to a lesser degree, the nucleus. Consistent with their localization patterns, we found that these helicases are required for ribosomal RNA processing and regulation of gene expression. Surprisingly, depletion of these helicases resulted in misregulation of highly overlapping sets of protein-coding genes, including many genes that function in differentiation and development. We conclude that R-loop-interacting DEAD-box helicases have non-redundant roles that are critical for maintaining the normal embryonic stem cell transcriptome.

Introduction

R-loops are nucleic acid structures that form when ssRNA invades into dsDNA to form RDHs, with ssDNA being looped out. Most of the R-loops form co-transcriptionally, in which nascent RNA transcribed by RNAP II threads back and anneals with the DNA it transcribed from. R-loops have been shown to regulate immunoglobulin class switching, DNA replication, transcription initiation and termination, and RNA modification (Yu *et al.*, 2003; Ginno *et al.*, 2012; Lombraña *et al.*, 2015; Marnef and Legube, 2020). Other than regulating genome in positive ways, R-loops have long been thought to contribute to genome instability, as the ssDNA of the R-loops can serve as a good target for DNA mutagens (Gómez-González and Aguilera, 2007), and stabilized R-loops in the genome may collide with DNA replication forks to induce chromosome rearrangements (Hamperl and Cimprich, 2016). Therefore, R-loops must be tightly regulated.

R-loops often have a relatively short half-life of 10-20 min (Sanz *et al.*, 2016). The factors that remove or inhibit R-loops formation may contribute to this short half-life, including topoisomerase Top1 (El Hage *et al.*, 2010), helicase DDX21 (Song *et al.*, 2017), nucleases RNase H1 and RNase H2 (Stein and Hausen, 1969; Cornelio *et al.*, 2017). To understand the factors that bind to and regulate R-loops in a systematic way, several groups have profiled the R-loop or RDH interactomes of human or mouse cell lines (Cristini *et al.*, 2018; Wang *et al.*, 2018; Li *et al.*, 2020). For studies that focus on endogenous R-loops, the S9.6 antibody has been used to enrich for RDHs from a pool of fragmented chromatin (Cristini *et al.*, 2018;

Li *et al.*, 2020). However, due to the fact that R-loop-proximal chromatin—encompassing hundreds of basepairs—is invariably co-purified with RDHs, it is difficult to distinguish R-loop-binding proteins from nearby chromatin proteins using these approaches. Although more aggressive DNA fragmentation may reduce the fraction of chromatin proteins co-purified with RDHs, RDHs are sensitive to extensive sonication and numerous non-sequence-specific nucleases. For studies that tried to capture RDH-interacting proteins *in vitro*, annealed RDHs were used to pull down their interacting proteins (Wang *et al.*, 2018). Compared to the S9.6-based *in vivo* assay, pull-down assay is able to identify proteins that specifically bind to RHD structures while excluding proteins binding to R-loop-proximal chromatin. However, it lacks proteins that bind to the ssDNA of R-loops, proteins that bind to R-loop structure as a whole, and since only a few oligos were used for pull-down experiments, it may enrich for factors that tend to bind the used sequences and deplete others. Therefore, a different strategy is required to profile R-loop-binding proteins.

In this chapter, we describe two proteomics approaches for more stringent identification of R-loop-associated proteins in mESCs. We identify overlapping sets of R-loop-interacting proteins and many new potential regulators of R-loops. Interestingly, we show that R-loop-binding proteins identified by these approaches are highly enriched in nucleolar proteins, consistent with the high levels of R-loops found within this compartment (El Hage *et al.*, 2010; Shen *et al.*, 2017; Velichko *et al.*, 2019). We find that several R-loop-binding helicase proteins appear to have

highly overlapping roles in processing of rRNA, as well as expression of coding genes, suggesting they function in a common pathway. Finally, we show that RNA-protein crosslinking traps a set of proteins that are lost in the absence of crosslinking due to transient or weak association with the RDHs. These studies provide a resource of stringent R-loop-associating proteins in mESCs, including multiple new potential regulators of R-loop formation or stability. In addition, our studies reveal that introduction of selective RNA-protein crosslinking can identify R-loop-binding proteins that are missed by standard approaches.

Results

Stringent capture of R-loop-binding proteins

To identify a stringent set of R-loop-associated proteins, we developed an S9.6 co-immunoprecipitation (co-IP) protocol that first uses high salt washes to remove most chromatin-associated proteins, followed by chromatin fragmentation and immunoprecipitation of RDHs (Fig. 2.1a). To validate this approach, we first tested whether S9.6 can pull down the RNAP II core subunit RPB1, RNA helicase DHX9, and the splicing factor SFPQ. DHX9 has been shown to bind to RDHs and R-loops *in vitro* (Chakraborty and Grosse, 2011), as well as associate with R-loops *in vivo*, where it regulates their accumulation (Chakraborty, Huang and Hiom, 2018; Cristini *et al.*, 2018). SFPQ is a splicing factor that was shown to reduce the accumulation of R-loops (Chakraborty, Huang and Hiom, 2018). We observed reproducible enrichment of these proteins (Fig. 2.1b), validating our approach. In addition, we verified that R-loops were enriched within S9.6 immunoprecipitates by quantitative PCR, observing higher signals at genes known to form R-loops (Chen *et al.*, 2015) in comparison to genomic regions that lack R-loops (Fig. 2.1c). As an additional control, we showed that addition of DNase I strongly reduced enrichment of co-immunoprecipitated proteins (Fig. 2.1d), while addition of RNase A strongly reduced enrichment of R-loops (Fig. 2.1e), further validating this approach. These studies demonstrate the high specificity with which R-loops and known interacting proteins are enriched by our approach.

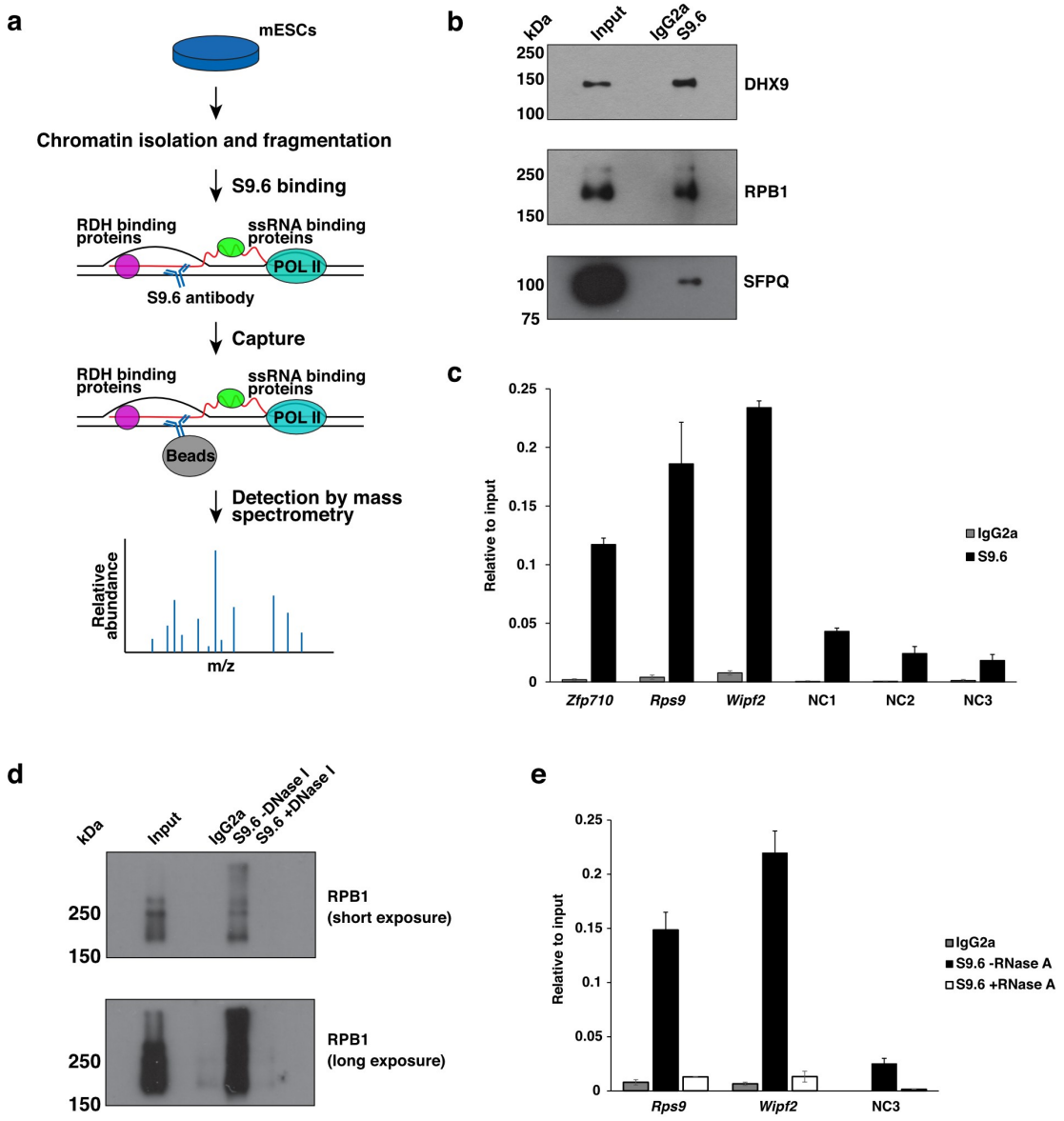

Figure 2.1. Validation of RDH and R-loop-binding proteins enrichment by S9.6 immunoprecipitation.

a. Schematic diagram of S9.6 co-IP experiment. **b.** Western blot validation of several proteins known to regulate R-loop formation. **c.** qPCR of DNA immunoprecipitated by S9.6. Primers were designed to target three genes known

to contain R-loops: *Zfp710*, *Rps9*, *Wipf2*, while NC1, NC2, NC3 are intergenic regions where no R-loops were previously detected. Enrichment is expressed relative to input DNA extracted from nuclear extract. **d.** Western blots of the S9.6 immunoprecipitates with and without DNase I treatment. **e.** qPCR of the S9.6 immunoprecipitates with and without RNase A treatment.

Identification of R-loop-binding proteins by mass spectrometry

Our lab previously mapped the genomic locations of R-loops in mESCs and showed that they regulate the recruitment of two chromatin remodeling complexes, making mESCs a good platform for studying R-loop functions (Chen *et al.*, 2015). To identify the R-loop interactome of mESCs in an unbiased manner, we performed S9.6 co-IP with or without RNase A treatment in three biological replicates. Samples were fractionated by SDS-PAGE (Fig. 2.2a) and subjected to tryptic digestion after isolating gel slices that excluded the majority of IgG heavy and light chains. Input and IP samples from each treatment were subjected to liquid chromatography-tandem mass spectrometry (LC-MS/MS) to identify the repertoire of proteins that interact with R-loops. A total of 709 proteins were detected in any of the three biological replicates in input or IP samples. As expected, very few proteins were identified in RNase A treated samples, demonstrating high specificity of the IP conditions (Fig. 2.2b). After filtering for proteins present in at least two IP replicates and removal of contaminating IgG peptides, 335 proteins were retained for downstream analysis (Table 2.1).

Next we used the Prostar software package to identify the most significantly enriched R-loop-interacting proteins (Wieczorek *et al.*, 2017). We observed 76 proteins that were enriched more than 2-fold (relative to normalized input) with p-value < 0.003 (Table 2.2, Fig. 2.2c). One of the most highly enriched proteins is DHX9, which was previously shown to interact with RDHs in HeLa cells, further validating our approach (Cristini *et al.*, 2018). Other highly enriched proteins

included SPB1 (gene name *Ftsj3*), a nucleolar protein that regulates pre-rRNA processing (Morello, Coltri, *et al.*, 2011) and NOG1 (gene name *Gtpbp4*), a nucleolar GTP-binding protein with crucial roles in 60S ribosome biogenesis (Jensen *et al.*, 2003). Another highly enriched protein was CHTOP, a component of the THO/TREX complex previously shown to inhibit the formation of R-loops (Huertas and Aguilera, 2003; Gómez-González *et al.*, 2011). In addition, numerous known or predicted RNA-binding proteins were identified, including DDX18, DDX21, DDX27, DDX54, which belong to the DEAD-box family of RNA helicases (Linder *et al.*, 1989; Cordin *et al.*, 2006; Linder, 2006). Another group of R-loop-interacting proteins, including PUM3 and HNRPU, have both DNA- and RNA-binding activities, suggesting possible roles in R-loop-dependent regulation of chromatin architecture or local epigenomic features. To validate our findings, we performed S9.6 co-IP followed by Western blotting and detected DDX18, DDX27, DDX54, with high IP efficiency (Fig. 2.2d). In all, 49% of proteins (164 of 335) identified in our S9.6 co-IP/MS analysis are known RNA-binding proteins, including several known regulators of R-loop formation, suggesting a high proportion of hits are bona fide R-loop-interacting proteins.

We next examined the enrichment of gene ontology (GO) categories among the 335 RDH-interacting proteins identified by our approach. As expected, multiple protein classifications associated with RNA processing were highly enriched (Fig. 2.3a). We observed similar enrichment by analyzing the 76 most prominent R-loop interacting proteins (enriched over two-fold, with $p < 0.003$, data not shown).

Notably, GO terms associated with ribosomal RNA processing were especially prominent, in agreement with the fact that R-loops are frequently observed in nucleoli (Shen *et al.*, 2017; Velichko *et al.*, 2019).

To examine the cellular localization of several stringent R-loop-interacting proteins, we performed immunofluorescence staining in mESCs. Consistent with the GO term analysis, DDX18, DDX24, and DDX27 showed substantial overlap with the nucleolar marker Fibrillarin (FBRL), although nuclear localization outside of the nucleolus was also observed, especially for DDX24 and DDX27 (Fig. 2.3b). Conversely, two less strongly enriched proteins, CTCF and SFPQ, exhibited largely diffuse nuclear staining (Fig. 2.3c). This raised the possibility that the strongest hits may be largely nucleolar, while hits with lower (but still significant) enrichment were more likely to interact with nuclear R-loops. Consistent with this possibility, R-loop-interacting proteins contributing to the enrichment of nucleolus- or rRNA-related GO terms exhibited stronger overall enrichment in our MS dataset (Fig. 2.3d). Finally, we compared our hits to a dataset describing the nucleolar proteome, which was previously measured by an independent group using isolated nucleoli from mouse fibroblasts followed by protein extraction and mass spectrometry quantification (Kar *et al.*, 2011). Of the 320 nucleolar proteins previously identified, we found that 122 overlapped with the 335 stringent R-loop-interacting proteins identified in our study (Fig. 2.3e, O/E: observed/expected = 22.8; p-value = 3.957×10^{-138}). Furthermore, of the 76 most prominently enriched R-loop-associated proteins (described above), 39 overlapped with the nucleolar

proteome (Fig. 2.3e, O/E = 32.1; p-value = 3.367×10^{-50}). These data indicate that many of the most strongly enriched R-loop-interacting proteins function largely within the nucleolus, likely due to the high abundance of co-transcriptional R-loops during rRNA synthesis. Accordingly, proteins with important roles in the regulation or functions of R-loops at protein coding genes may be enriched to lower levels in the LC-MS/MS dataset.

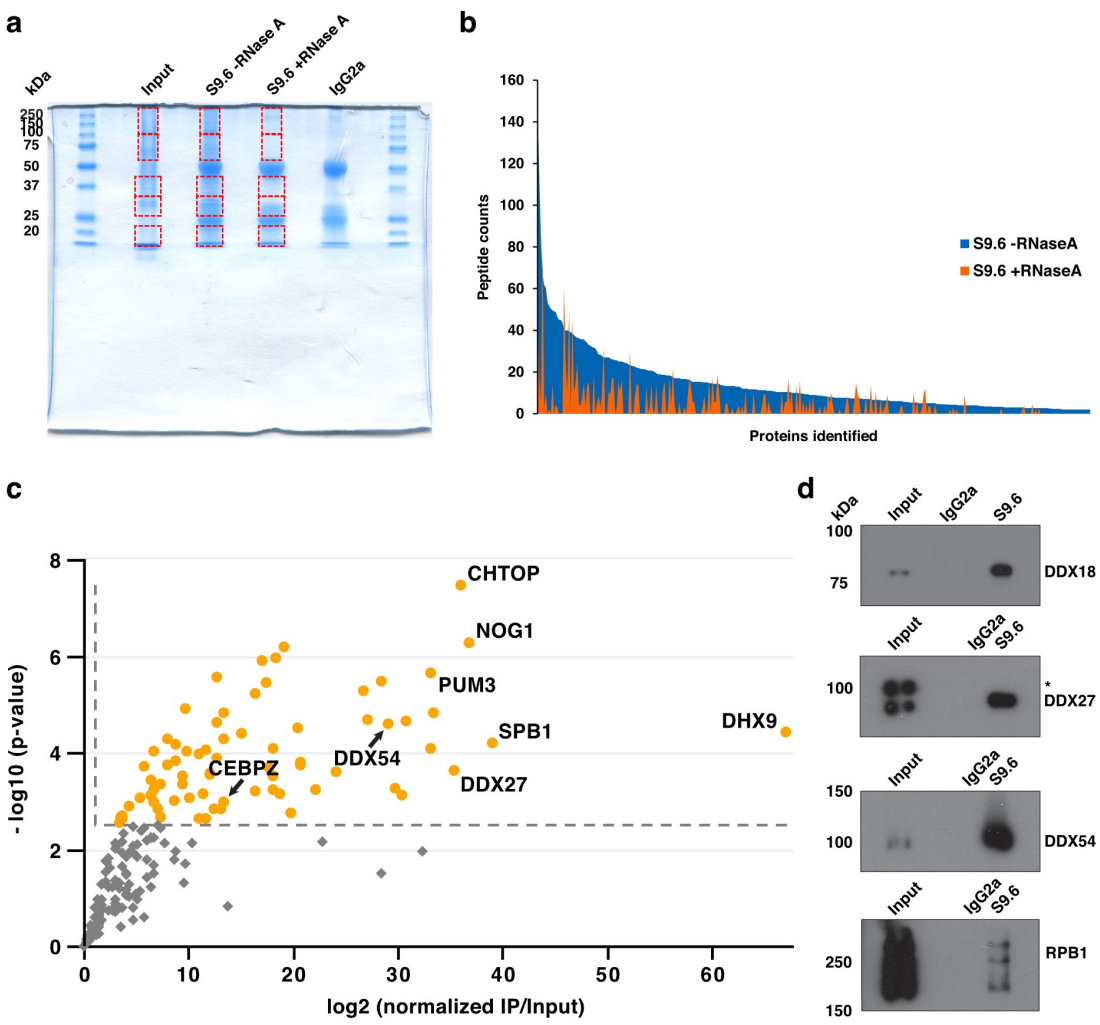

Figure 2.2 Identification of R-loop-binding proteins.

a. Coomassie blue staining of S9.6 immunoprecipates separated by SDS-PAGE. Gel slices were isolated (dashed squares) and sent for MS. The gel corresponding to the first biological replicate of three is shown. **b.** Peptide counts of proteins identified with or without RNase A treatment. Peptide counts were averaged from three biological replicates. Proteins identified in 2 or 3 replicates in -RNase A experiments are shown in blue, sorted by peptide counts. Average peptide counts

of the corresponding proteins from +RNase A experiments are shown in orange**.** **c.** Volcano plot of the proteins enriched by S9.6 IP. The vertical dashed line denotes two-fold normalized enrichment and the horizontal dashed line denotes a p-value of 0.003 [-log10 (p-value) = 2.5]. 76 highly enriched proteins are shown as orange dots. Highly enriched proteins of interest were labeled with their protein name. Three biological replicates were included. **d.** Western blots of several DEAD-box family proteins enriched by S9.6 co-IP. mIgG2a is a negative control for IP and the RNAP II subunit RPB1 is shown for comparison. The asterisk denotes a non-specific band.

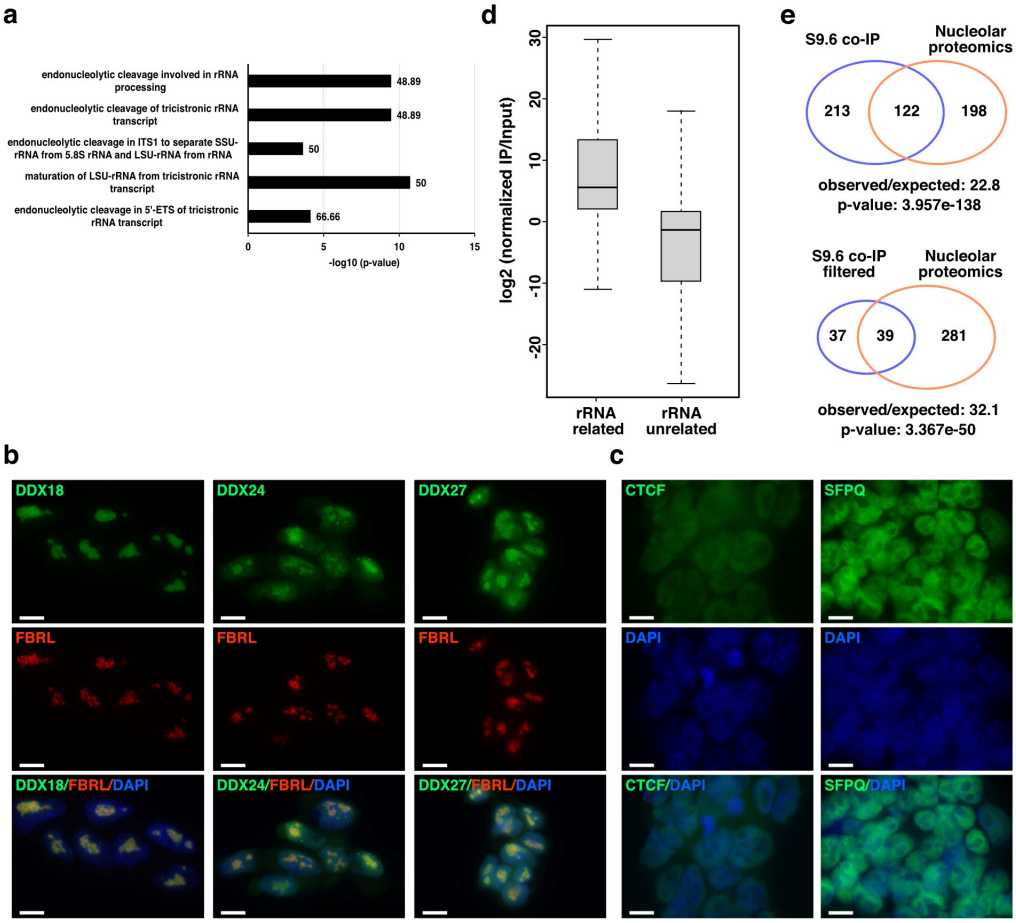

Figure 2.3. Many R-loop-binding proteins are enriched within the nucleolus.

a. GO term analysis of R-loop-interacting proteins identified by MS. PANTHER GO term analysis was performed for “GO biological process” categories. The top five pathways are shown. **b.** Immunofluorescence staining of three R-loop-associated DEAD-box proteins, DDX18, DDX24 and DDX27, co-stained with FBRL to mark the nucleolus. DNA was stained by DAPI. Scale bar = 10 μm. **c.** Immunofluorescence staining of two proteins, CTCF and SFPQ, with relatively low

levels of enrichment within S9.6 immunoprecipitates. **d.** Comparison of rRNA-related and rRNA-unrelated proteins enriched by S9.6. R-loop-associated proteins were classified as rRNA related or unrelated by GO term analysis, using the biological process term “ribosomebiogensis”, “rRNAmetabolicprocess”, “rRNAprocessing”. Boxes of the box plot represent the first and third quartiles, the band represents the median, and the whiskers depict 1.5 times the interquartile range. **e.** Venn diagram of the overlap between the R-loop interactomes (with and without filtering) and the nucleolar proteome from Kar *et al.* (Kar *et al.*, 2011). The upper diagram includes the 335 proteins enriched in S9.6 co-IP samples with the 76 most enriched (corresponding to the orange dots in Fig. 2.2c) depicted in the lower diagram. P-values were calculated using hypergeometric tests.

Table 2.1. List of 335 proteins present in at least two IP replicates

Dhx9	Surf6	Rps18	Ddx10	Mcm5	Snrnp40
Ftsj3	Rpl4	Wdr3	Rpl24	Cdc5l	Rpn1
Gtpbp4	Wdr74	Ccdc137	Prpf19	Snrpe	Chd4
Chtop	Rpl21	Rps24	Rpl28	Xrn2	H2ax
Ddx27	Utp4	Rrp12	Baz2a	Lbr	Ddx17
Bop1	Hnrnpul2	Aatf	Rps23	Nop9	Phb
Nat10	Ccdc59	C1orf13 homolog	Snrpd3	D1Pas1	Ptbp1
Pum3	Nol7	Rps11	Sap18	Slc25a4	Eftud2
Utp20	Tbl3	Rpl34	Dimt1	Utf1	Elavl1
Nop2	Fam207a	Rpl37a	Abt1	C1qbp	Ppp1cc
Rpl6	Rpl32	Rpl14	Nup160	Srsf5	Ppp1cb
Utp14a	Rps19	Rpl13	Ssr1	Ddx3y	Hnrnpab
Ddx54	Nop56	Rpl15	Fyttd1	Ywhaq	H2aw
Mybbp1a	Imp4	Rrp8	H1-2	Alyref	Vdac1
Rbm28	Rpl30	Nip7	H1-6	Slc2a1	Sf3b1
Hnrnpu	Rpl8	Rps14	Nup155	Eif6	Rps27a
Rpl5	Rpl35	Llph	Gar1	Ywhag	Hsp90ab1
Rpf2	Rpl35a	Rps8	Rpl9	Rps2	Hsp90aa1
Ddx21	Rps9	Rpl12	Dppa2	Trip12	Vim
Prpf8	Dkc1	Rpl11	Rps20	Snrnp70	Kpna2
Gnl3	Rbm14	Utp3	Rbmxl1	U2af1	H2bc1
Pdcd11	Nsa2	Utp11	Noc2l	Ddx3x	Ppp1ca
Wdr46	Rrp7a	Dcaf13	Dhx15	Srsf3	Nup93
Hnrnpm	Pwp1	Mov10	Tra2a	Nup107	Acta2
Nifk	Rpl27a	Rplp0	Myef2	Rps3	Actg2
Noc3l	Nop58	Cenpv	Rplp2	Rpsa	Hnrnpa3
Pes1	Rpl23	Fcf1	Slc25a13	Bclaf1	Actc1
Rps4x	Rps15	Nhp2	Nup133	Rbm39	Acta1
Utp6	Rps19bp1	Srsf10	Rps10	Esrrb	Phb2
Brix1	Rpl18	Nol10	Cmss1	Ywhaz	H2bu1
Ddx18	Hnrnpf	Rrp15	Rps26	Atp5f1c	Hist3h2ba
Ppan	Rcl1	Kri1	Fbl	Gnai2	Smarca5
Rpl7a	Mphosph10	Nol9	Snrpd2	Ppia	H2bc21
Rsl1d1	Alb	Glyr1	Pnn	Hnrnpdl	Srsf1
Utp15	Gnl2	Rrp36	Ctcf	Srsf7	Lmnb2
Utp18	Rpl7l1	Tra2b	Rps17	Pcbp2	Top2a

Rrs1	Ddx56	Rps5	Rbm19	Slc25a3	Hnrnpk
Rpl7	Rsl24d1	Nol12	Dppa4	Macroh2a1	Canx
Wdr43	Mrto4	Rpl36	Srsf6	Slc25a5	H2bc3
Cebpz	Rpl37	C11orf98 homolog	H1-1	Cbx1	H2bc12
Rpl23a	Poldip3	Rpl31	Atp5po	Hnrnpa0	H2bc9
Rpl3	Erh	H1-5	Gapdh	Pcbp1	Hnrnpl
Raly	Dnttip2	H1-3	Rps7	Snrpa	Hnrnpa2b1
Rpl17	Isg20l2	H1-4	Tomm22	Rack1	Trim28
Nol11	Matr3	Imp3	Rpl22	Ddx5	Kpnb1
Rrp1b	Noc4l	Rpl27	Srsf4	Vdac2	Actb
Ebna1bp2	Nop53	Tuba1b	Lin28a	Rae1	Hnrnpa1
Rrp1	Wdr75	Rpl10	Rps3a	H2az1	H3c2
Nop14	Mak16	Rpl13a	Pno1	H2az2	H3c1
Rrp9	Metap1	Ilf3	Dnmt3b	Cbx3	Actg1
Rps13	Rpl29	Cdca8	Snu13	Ywhae	Hspa8
Rpl18a	Rpl36a	Ngdn	Sec61b	Nvl	Npm1
Ddx24	Nop16	Rps25	Rps15a	Tardbp	Lmnb1
Hnrnpc	Nol6	Rplp1	Ilf2	Pabpc1	H4c1
Rpl26	Rps6	Ddx52	Ywhab	Nup98	Ncl
Pwp2	Rps16	Utp23	Rps12	Hnrnpd	

Table 2.2. List of 76 highly enriched proteins

Dhx9	Hnrnpu	Brix1	Rpl17	Rpl21	Rpl27a
Ftsj3	Rpl5	Ddx18	Nol11	Utp4	Rps15
Gtpbp4	Rpf2	Ppan	Rrp1b	Ccdc59	Rps19bp1
Chtop	Prpf8	Rpl7a	Ebna1bp2	Nol7	Rpl18
Ddx27	Gnl3	Rsl1d1	Rrp1	Tbl3	Gnl2
Bop1	Pdcd11	Utp15	Nop14	Rpl32	Rpl37
Nat10	Wdr46	Utp18	Rrp9	Rps19	Rpl29
Pum3	Hnrnpm	Rpl7	Rps13	Imp4	Rpl34
Nop2	Nifk	Wdr43	Rpl18a	Rpl30	Rpl37a
Rpl6	Noc3l	Cebpz	Ddx24	Rpl8	Rpl14
Utp14a	Pes1	Rpl23a	Rpl26	Rpl35a	Rrp8
Ddx54	Rps4x	Rpl3	Pwp2	Rbm14	
Rbm28	Utp6	Raly	Surf6	Pwp1	

Regulation of rRNA processing by R-loop-interacting DEAD-box family helicases

The DEAD-box protein family, named after the conserved D-E-A-D amino acid sequence within the Walker B motif, consists of known or putative ATP-dependent RNA helicases that are conserved throughout eukaryotes (Linder *et al.*, 1989; Cordin *et al.*, 2006; Linder, 2006). DEAD-box proteins have been found to contribute to RNA metabolism *in vivo*, including processes such as mRNA transcription and degradation, splicing, mRNA export, and ribosome biogenesis (Linder, 2006).

Of the 335 stringent R-loop-associated proteins we identified from our MS, 13 belong to the DEAD-box protein family, including known R-loop regulators such as DDX5 (Villarreal *et al.*, 2020) and DDX21 (Song *et al.*, 2017). To better understand the functions of this protein family in the regulation of R-loops, we used RNA interference to examine the cellular consequences of partial depletion of R-loop-associated proteins DDX10, DDX24, DDX27, DDX54. Several DEAD-box family members were previously shown to localize to the nucleolus and function in ribosome biogenesis (Zagulski *et al.*, 2003; Turner *et al.*, 2009; El Hage *et al.*, 2010; Srivastava *et al.*, 2010; Saporita *et al.*, 2011). For example, DDX5 has been shown to promote rRNA transcription (Saporita *et al.*, 2011), whereas DDX51 and DDX54 were shown to affect different steps of rRNA processing (Srivastava *et al.*, 2010; Milek *et al.*, 2017). We therefore examined the roles of DDX10, DDX24, DDX27, and DDX54 in the production and processing of rRNAs. To this end, we

performed Northern blotting on total RNA isolated from mESCs, using probes specific to the 18S or 28S rRNAs. In addition to fully processed rRNAs, these probes also hybridize to the 45S pre-rRNA and multiple smaller rRNA processing intermediates (Morello, Hesling, *et al.*, 2011; Henras *et al.*, 2015; Moraleva *et al.*, 2017), as outlined in Fig. 2.4a. The 18S rRNA Northern probe, which also detects the 45S, 41S, and 34S pre-rRNAs, uncovered alterations in levels of 45S and 34S pre-rRNAs in *Ddx24* and *Ddx10* KD, respectively (Fig. 2.4b). In contrast, the 28S probe revealed increased 36S pre-rRNA in *Ddx24* KD and *Ddx27* KD mESCs (Fig. 2.4c). In addition, *Ddx54* KD cells exhibited increased 32S pre-rRNA relative to *EGFP* KD control cells (Fig. 2.4c). Collectively, these data demonstrate that DDX10, DDX24, DDX27, and DDX54 contribute to production of mature 18S and 28S rRNAs at the level of rRNA processing, as previously observed for DEAD-box helicases DDX5, DDX17, and DDX51 (Jalal, Uhlmann-Schiffler and Stahl, 2007; Srivastava *et al.*, 2010; Saporita *et al.*, 2011).

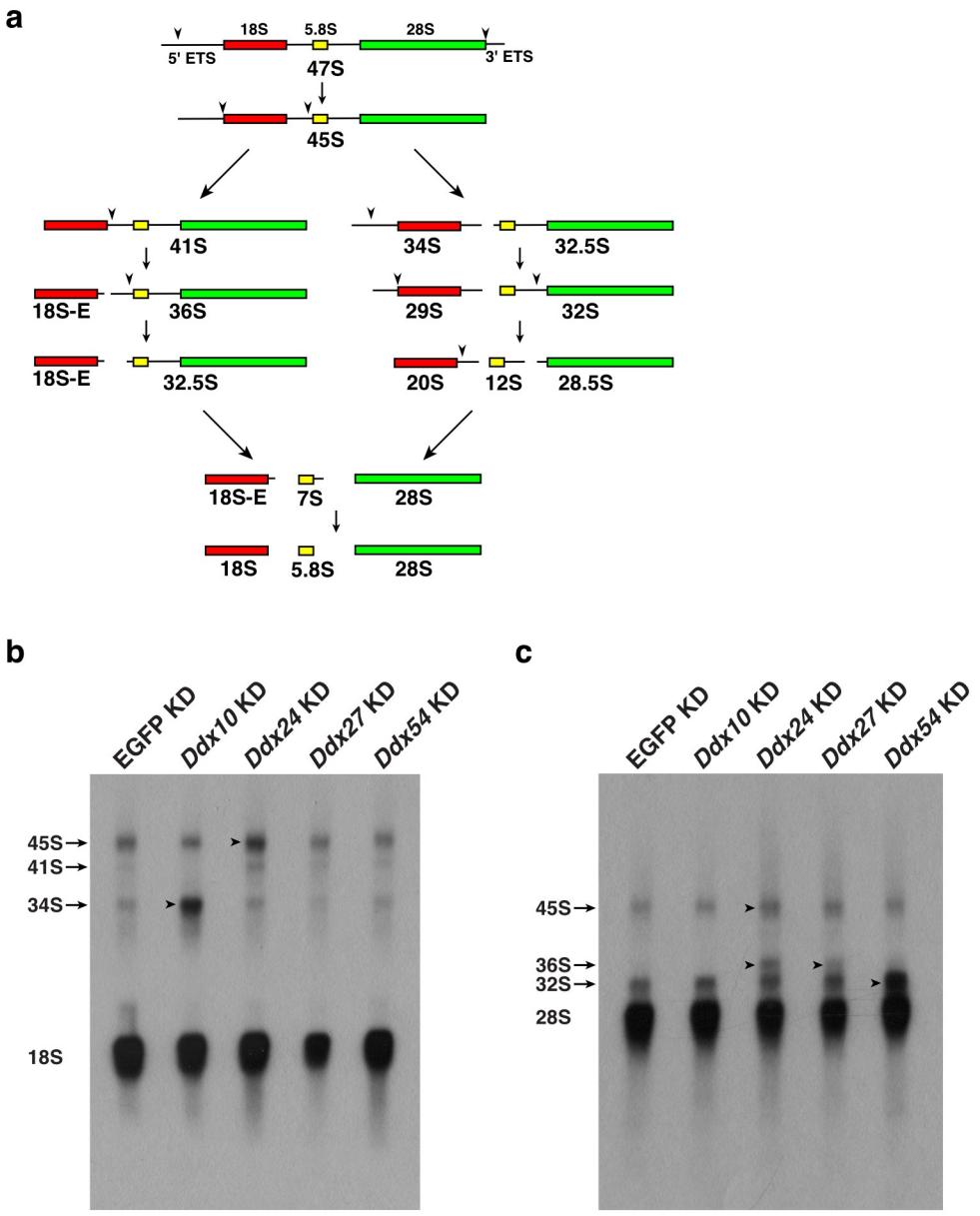

Figure 2.4. Regulation of rRNA processing by R-loop-interacting DEAD-box family helicases.

a. Schematic diagram of rRNA processing in mouse. **b.** Relative abundance of 18S rRNA and indicated pre-rRNA transcripts in EGFP and *Ddx* KDs, as shown in a Northern blot using a probe targeting the 18S sequence. Arrowheads indicate alterations in pre-mRNA transcripts in some KDs. **c.** Northern blot using a probe targeting the 28S sequence, as depicted in b.

Shared functions of DEAD-box proteins in regulation of mRNA expression

Some of the DEAD-box proteins have not only been found regulate rRNA transcription and processing, but also been shown to regulate mRNA transcription. For example, DDX17 and DDX24 have been shown to regulate rRNA processing as well as transcription of genes regulated by estrogen receptor-alpha (Wortham *et al.*, 2009; Song *et al.*, 2017). DDX5, which has been shown to promote rRNA transcription, also interacts with zinc finger protein PLZF and binds at its binding sites, such as promoter of *ILF3* gene and promotes its transcription (Saporita *et al.*, 2011; Legrand *et al.*, 2019).

Based on IF, DDX24 and DDX27 exhibited strong nucleolar localization but also some localization within non-nucleolar regions of the nucleus (Fig. 2.3b). To test these proteins for potential roles in expression of protein-coding genes, we performed mRNA-seq upon knockdown of the same set of DEAD-box proteins. Using a cutoff of 2-fold up- or down-regulated in any knockdown relative to control and adjusted p-value < 0.05, we observed 737 genes that were differentially expressed, with more genes up-regulated than down-regulated, in one or more *Ddx* KD relative to controls (Fig. 2.5a). Consistent with the high correlation of their overall expression profiles, the sets of genes up-regulated or down-regulated by depletion of each factor were highly overlapping. Of the 603 genes significantly up-regulated in any of the four KDs, 209 were shared among all four of the KDs (Fig. 2.5b), including transcripts such as *Cdkn1a*, *Cdkn2b*, and *Wnt5a*. The observed changes in expression of these transcripts upon knockdown of each DEAD-box

protein were further confirmed by RT-qPCR (Fig. 2.5f-h). Of the 134 genes significantly downregulated in at least one knockdown, 19 were downregulated in all four KDs (Fig. 2.5c). This overlap was also significantly more than expected (p-value = 1.27×10^{-133}), albeit lower than observed for upregulated genes. The substantial overlap of genes misregulated by knockdown of each factor suggested either that DDX10, DDX24, DDX27, and DDX54 regulate mRNA levels through a shared pathway, or that the genes observed to be misregulated were more sensitive to the alterations in rRNA levels observed upon knockdown of these proteins. Notably, GO term analysis of upregulated genes revealed enrichment for multiple terms related to cellular differentiation, including “neuronal differentiation”, “cell fate commitment”, and “cell migration” (Fig. 2.5d). Similarly, down-regulated genes were also enriched for terms related to development and differentiation, including genes that regulate lipid and lipoprotein homeostasis, such as *Adipoq* (Hu, Liang and Spiegelman, 1996) and *Pcsk9* (Wu and Li, 2014) (Fig. 2.5e, 2.5i). These findings raise the possibility that some alterations in gene expression upon depletion of these DEAD-box family helicases may be specific to ESCs, with potential implications for maintenance of the pluripotent state or regulation of ESC differentiation.

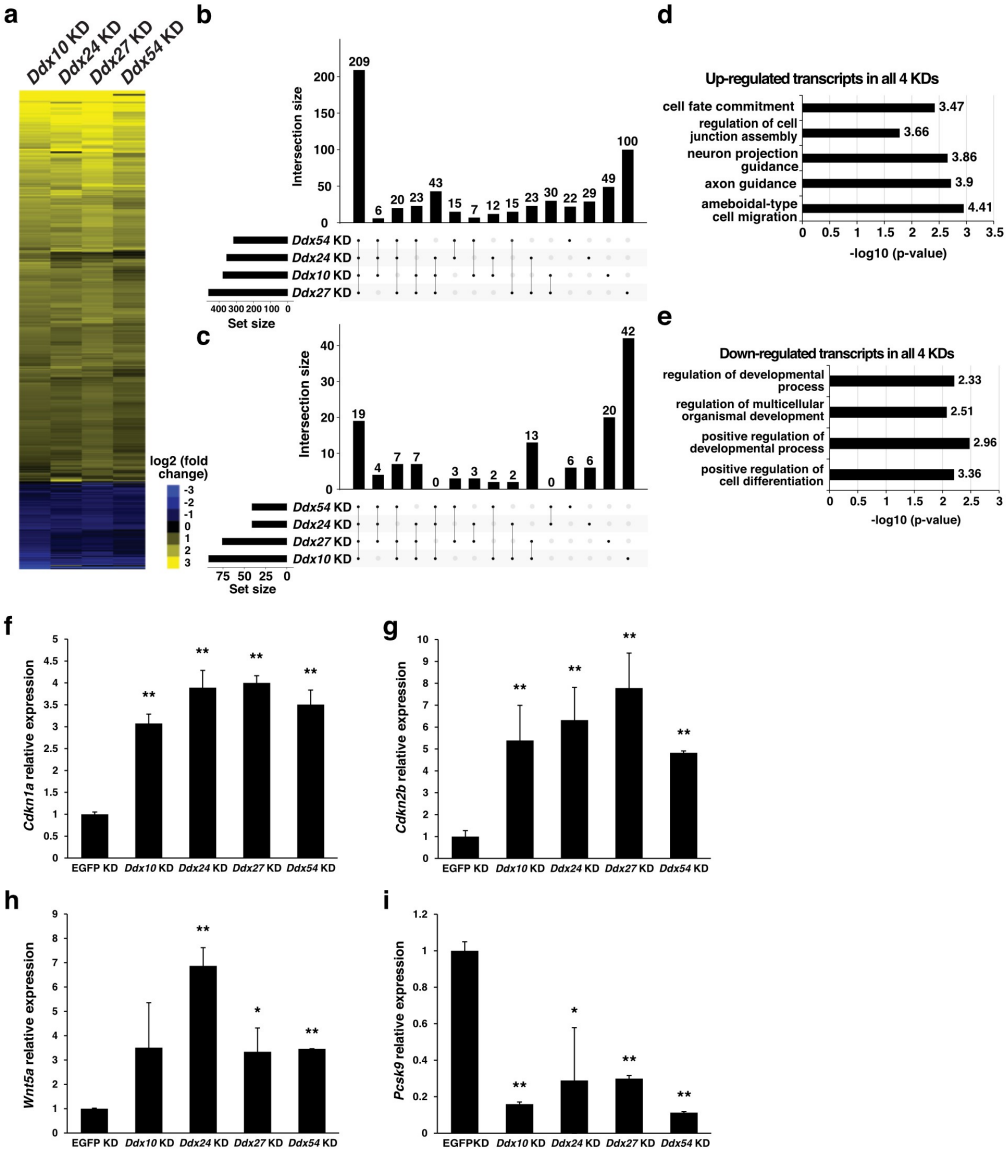
a
Ddx10 KD
Ddx24 KD
Ddx27 KD
Ddx54 KD
log2 (fold change)
-3
-2
-1
0
1
2
3
b
Intersection size
209
6
20
23
43
15
7
12
15
23
30
22
29
49
100
Ddx54 KD
Ddx24 KD
Ddx10 KD
Ddx27 KD
400 300 200 100 0
Set size
c
Intersection size
19
4
7
7
0
3
3
2
2
13
0
6
6
20
42
Ddx54 KD
Ddx24 KD
Ddx27 KD
Ddx10 KD
75 50 25 0
Set size
d
Up-regulated transcripts in all 4 KDs
cell fate commitment
3.47
regulation of cell junction assembly
3.66
neuron projection guidance
3.86
axon guidance
3.9
ameboidal-type cell migration
4.41
-log10 (p-value)
e
Down-regulated transcripts in all 4 KDs
regulation of developmental process
2.33
regulation of multicellular organismal development
2.51
positive regulation of developmental process
2.96
positive regulation of cell differentiation
3.36
-log10 (p-value)
f
Cdkn1a relative expression
EGFP KD
Ddx10 KD
Ddx24 KD
Ddx27 KD
Ddx54 KD
g
Cdkn2b relative expression
EGFP KD
Ddx10 KD
Ddx24 KD
Ddx27 KD
Ddx54 KD
h
Wnt5a relative expression
EGFP KD
Ddx10 KD
Ddx24 KD
Ddx27 KD
Ddx54 KD
i
Pcsk9 relative expression
EGFPKD
Ddx10 KD
Ddx24 KD
Ddx27 KD
Ddx54 KD

Figure 2.5. Regulation of mRNA expression by R-loop-interacting DEAD-box proteins.

a. Genes significantly misregulated by *Ddx* knockdown as measured by mRNA-seq. The 737 transcripts at least 2-fold up- or down-regulated in any knockdown relative to control (with a p-value < 0.05) are depicted in the heatmap. **b.** Overlap of genes significantly up-regulated in any of the four KDs. Linked dots represent genes shared between the indicated KDs. **c.** Overlap of the genes significantly down-regulated in any of the four KDs, depicted as in b. **d.** GO term analysis of the genes significantly up-regulated in all four KDs. The top five GO pathways are shown. **e.** GO term analysis of the genes significantly down-regulated. The only four pathways reaching statistical significance are shown. **f.** RT-qPCR confirmation of the *Cdkn1a* gene shown by mRNA-seq to be up-regulated in all *Ddx* KDs. Expression levels were normalized to *Gapdh* levels. Three technical replicates were included. **g.** Expression of the *Cdkn2b* gene, depicted as in f. **h.** Expression of the *Wnt5a* gene, depicted as in f. **i.** Confirmation of the *Pcsk9* gene shown by mRNA-seq to be down-regulated in four KDs.

RNA-protein crosslinking uncovers transiently or weakly interacting R-loop-associated proteins

Although proteins that bind to or near R-loops can be identified by co-immunoprecipitation using the S9.6 antibody, it is difficult to distinguish proteins that bind directly to RDHs from the proteins that bind to chromatin flanking the RDHs. In addition, while our stringent chromatin wash reduces background in our co-immunoprecipitation procedure, it likely also removes RDH-associated proteins that interact transiently or weakly with these structures. To systematically and specifically identify proteins that directly bind to the RDH structure of R-loop, we adapted a crosslinking protocol previously used to enrich for RNA-associated proteins (He *et al.*, 2016) and coupled this procedure with S9.6 co-immunoprecipitation.

To this end, we cultured mESCs with 4-thiouridine (4SU) to incorporate this nucleotide analog into RNA transcripts, followed by irradiation with intermediate wavelength (312 nm) UV light, which has been shown to induce crosslinks between 4SU-labeled RNA and their direct binding proteins (Hafner *et al.*, 2010; He *et al.*, 2016). To precisely quantify the extent to which crosslinking affected enrichment of proteins, we introduced stable isotope labeling by amino acids in cell culture (SILAC) (Ong *et al.*, 2002; Mann, 2006) to directly compare protein abundance in the presence or absence of 4SU addition. After 4SU or vehicle addition and UV treatment, cells were lysed and their chromatin was mixed in equal

amounts, followed by S9.6 IP and LC-MS/MS for each of three independent replicates (Fig. 2.6a).

After removing proteins that appeared only in one replicate, 231 proteins were left for downstream analysis (Table 2.3). Interestingly, although we observed enrichment of 116 proteins upon 4SU addition, another group of 115 proteins was reduced in the 4SU samples relative to non-4SU-containing samples. Whereas the enriched proteins were likely proteins that directly interact with the RNA part of R-loops, either transiently or weakly, those that were reduced in the crosslinked samples may represent proteins that interact very strongly with R-loops with multiple RNA contacts throughout each polypeptide. For proteins in this latter category, multiple protein-RNA crosslinks may render a substantial fraction of peptides “unreadable” by mass spectrometry, due to the covalent attachment of oligonucleotides of unknown mass that remain even after bulk digestion of RNA in the immunoprecipitates. Although 202 of 231 proteins identified in the crosslinking dataset overlapped with the 335 R-loop-interacting proteins identified in our initial studies, an additional 29 proteins were recovered in the crosslinked samples, for a total of 364 R-loop-binding proteins overall (Table 2.3). Based on GO term analysis, these 29 additional proteins included proteins that function in mRNA splicing or transport, raising the possibility that some of these proteins were only weakly or transiently associated with R-loops due to active roles in removing transcripts from contexts conducive to R-loop formation.

To explore the possibility that crosslinking may enrich weakly interacting proteins while depleting strongly interacting proteins, we directly compared the two datasets, hereafter referred to as “crosslinked” and “uncrosslinked”. For this comparison, we included proteins that appeared in two or more replicates in the uncrosslinked data, and further filtered this list to include only those proteins identified by both mass spectrometry approaches, leaving 214 total proteins. Interestingly, we found that the proteins with higher enrichment in the presence of 4SU than in its absence were among the most poorly enriched in the uncrosslinked data relative to input (Fig. 2.6b, upper left quadrant). Conversely, the proteins for which enrichment was reduced by crosslinking were often among the highly enriched in the uncrosslinked data (Fig. 2.6b, lower right quadrant). Consequently, we observed a negative correlation between the two datasets (Fig. 2.6b). To further validate these results, we performed S9.6 co-IP followed by quantitative Western blotting, with or without crosslinking, on a sample of proteins across the enrichment spectrum. We examined three proteins (CEBPZ, DDX18, and DDX27) reduced in the presence of 4SU-dependent crosslinking and one (SFPQ) enriched in the presence of crosslinking (Fig. 2.6b). Consistent with the mass spectrometry data, the quantitative Western blotting showed an anti-correlation pattern (Fig. 2.6c). Overall, these findings suggest crosslinking improves enrichment of weakly or transiently RDH-interacting proteins that bind directly to RNA but reduces enrichment of some strongly associated proteins. Previous studies of RNA-protein interactions revealed that proteins and protein domains directly associated with

RNA can be underrepresented in the spectra upon crosslinking, due to the covalent attachment of RNA nucleotides (He *et al.*, 2016). A similar phenomenon may explain the crosslinking-dependent reduction we observe for strongly enriched R-loop-associated factors.

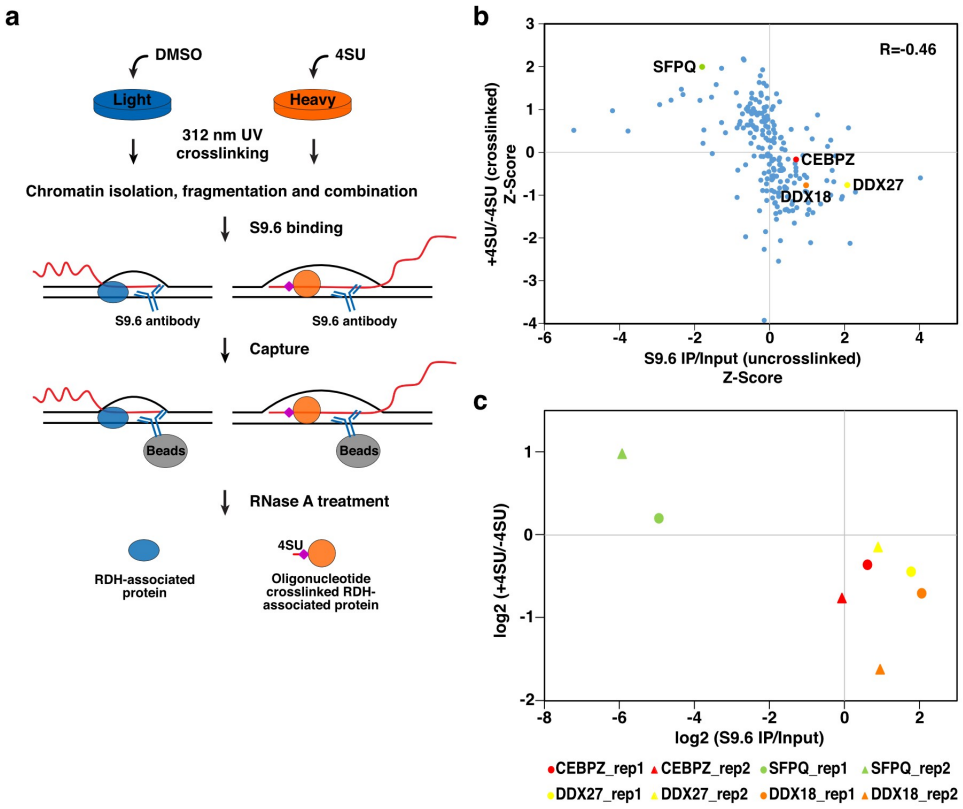

Figure 2.6. RNA-protein crosslinking enables identification of weakly-interacting R-loop-binding proteins.

a. Schematic diagram of 4SU-labeled crosslinking R-loop purification approach. **b.** Comparison of protein enrichment by S9.6 IP in the presence or absence of crosslinking. Normalized enrichment of the 214 proteins identified by both methods are plotted, and the Pearson's correlation coefficient is indicated. **c.** Measurement of the effect of crosslinking on selected R-loop-interacting proteins by quantitative Western blotting. Two biological replicates were performed.

Table 2.3. List of proteins identified in uncrosslinked and crosslinked IP

202 proteins shared between uncrossslinked and crosslinked			133 proteins specific to uncrosslinked		29 proteins specific to crosslinked
Atp5po	Rps5	Utp18	Nat10	Dnmt3b	Srsf2
Xrn2	H2bc1	Ngdn	Utp20	Snu13	Fus
Srsf5	Tra2b	Rpl12	Prpf8	Sec61b	Sfpq
Rps10	Rpl29	Cdc5l	Pdcd11	Ywhab	Nono
Srsf1	H1-1	Rpl5	Rsl1d1	Rps12	Strbp
Rpl24	Rps9	Rpl4	Nol11	Mcm5	Npm3
Hnrnpa1	Rbm39	Rrp7a	Rrp1b	Snrpe	Ubb
Hnrnpab	Dkc1	Utp14a	Rrp9	Lbr	Rbmx
Srsf7	Gar1	Rpl17	Ddx24	Nop9	H2ac11
Erh	Rps18	Rpl28	Pwp2	Slc25a4	Snrpa1
Actb	Wdr46	Ddx27	Utp4	C1qbp	Csnk2a1
Ilf2	Mybbp1a	Ddx18	Tbl3	Ywhaq	Igf2bp1
Hnrnpk	Rpl27	Fam207a	Rpl35a	Slc2a1	Syncrip
Pcbp2	Rplp0	Rpl18	Rps19bp1	Ywhag	U2af2
Ilf3	Rps6	Rpl21	Alb	Trip12	Khsrp
Srsf3	Rps19	Hnrnpul2	Rpl37	Nup107	H2bc7
Hspa8	Hnrnpl	Fcf1	Poldip3	Rpsa	H2bc14
Hnrnpa2b1	Macroh2a1	Rbm19	Noc4l	Bclaf1	Hist2h2bb
Rps15	Hnrnpu	Rpl13a	Nop53	Atp5f1c	H2bc4
Utf1	Alyref	Ppan	Wdr75	Gnai2	Khdrbs1
Hnrnpa0	Nhp2	Nol6	Metap1	Ppia	Hnrnph2
Npm1	Lin28a	Brix1	Rpl36a	Slc25a3	Hnrnph1
Ddx17	Rpl35	Ftsj3	Wdr3	Cbx1	H3-5
Rpl14	Nol7	Ddx56	Ccdc137	Snrpa	Rpl10a
Snrpd3	Top2a	Rcl1	Rrp12	Rack1	Rbm8a
H2az1	Imp3	Rbm28	C1orf13 homolog	Vdac2	Krr1
H2az2	Cdca8	Rsl24d1	Rpl15	Rae1	Ssb
Fyttd1	Utp15	Bop1	Rrp8	Pabpc1	Rpf1
Ywhaz	Rpl6	Surf6	Llph	Nup98	Psmd1
Ywhae	Esrrb	Elavl1	Dcaf13	Snrnp40	
Rps8	Rpl31	Nip7	Mov10	Rpn1	
H4c1	Srsf10	Nop2	Cenpv	Chd4	
Ddx5	Fbl	Pum3	Nol10	H2ax	
Rbmxl1	Hnrnpf	Utp23	Rrp15	Phb	

Pcbp1	Rpl26	Ebna1bp2	Nol9	Ppp1cb	
Rps3	Rps24	Mak16	Glyr1	H2aw	
D1Pas1	Utp11	Noc3l	Rrp36	Vdac1	
Ddx3x	Mphosph10	Nop16	Nol12	Sf3b1	
Ddx3y	Cebpz	Rpf2	C11orf98 homolog	Rps27a	
Rps3a	Utp3	Nop14	Tuba1b	Hsp90ab1	
Hnrnpd	Rpl3	Mrto4	Rpl10	Hsp90aa1	
Hnrnpdl	Kri1	Rrs1	Rplp1	Vim	
Ddx21	Pwp1	Wdr74	Ddx52	Kpna2	
Rps2	Rpl23	Nifk	Ddx10	Ppp1ca	
Rps25	Tardbp	Ccdc59	Prpf19	Nup93	
Rps17	Wdr43	Rrp1	Baz2a	Acta2	
H1-5	Aatf	Eif6	Rps23	Actg2	
Slc25a5	Nop56	Hnrnpc	Sap18	Actc1	
H1-6	Rpl30	Rpl7l1	Dimt1	Acta1	
Rps15a	Rpl8	Gnl3	Abt1	Phb2	
Rps14	Rpl32	Matr3	Nup160	H2bu1	
Tra2a	Rpl13	Pes1	Ssr1	Hist3h2ba	
Hnrnpa3	Rpl27a	Isg20l2	Nup155	Smarca5	
Ptbp1	Eftud2	Raly	Dppa2	H2bc21	
Cbx3	Ddx54	Rpl9	Rps20	Lmnb2	
Rps16	Imp4	Nvl	Rplp2	Canx	
H1-2	Nop58	Nsa2	Slc25a13	H2bc3	
H1-4	Rbm14	Hnrnpm	Nup133	H2bc12	
H1-3	Rpl7	Gtpbp4	Cmss1	Trim28	
U2af1	Rpl36	Myef2	Rps26	Kpnb1	
Chtop	Noc2l	Gnl2	Snrpd2	H3c2	
Ppp1cc	Rpl7a	Dhx15	Pnn	H3c1	
Rps4x	Rpl23a		Ctcf	Actg1	
Rps13	Rpl34		Dppa4		
Rps11	Rpl11		Srsf6		
Snrnp70	Rpl18a		Gapdh		
H2bc9	Dnttip2		Tomm22		
Ncl	Utp6		Rpl22		
Rps7	Dhx9		Srsf4		
Lmnb1	Rpl37a		Pno1		

Discussion

In this chapter, we established approaches for identification of R-loop-associated proteins, using a modified immunoprecipitation approach that includes stringent extraction of chromatin proteins prior to pull-down, with or without RNA-protein crosslinking. These approaches uncovered several hundred associated proteins, including a high portion of RNA-interacting proteins and several known R-loop regulators. One of the most enriched proteins was DHX9, which has previously been shown to bind R-loops and regulate their formation (Chakraborty, Huang and Hiom, 2018; Cristini *et al.*, 2018). Moreover, DHX9 is also known to bind G-quadruplexes (G4s) within DNA (Chakraborty and Grosse, 2011), suggesting a possible role for DHX9 in coordinate regulation of R-loops and G4s where they colocalize in the genome (Miglietta, Russo and Capranico, 2020).

Upon examining the functions of R-loop interacting proteins identified in this study, GO terms related to rRNA were enriched, and factors that localize within the nucleolus and function in rRNA processing were particularly prominent. The proteins most highly enriched by S9.6 tend to be nucleolar, often with known or predicted functions in rRNA production or processing. These findings lend additional support to the idea that a significant fraction of R-loops within cells is located within nucleolus due to RNAP I mediated transcription (El Hage *et al.*, 2010; Shen *et al.*, 2017).

Members of the DEAD-box family of RNA helicases were overrepresented among the R-loop-associated RNA-binding proteins identified. DEAD-box helicases are known to be involved in RNA processing, including transcription, splicing, and RNA decay (Rocak and Linder, 2004). Several DEAD-box proteins have established roles in the regulation of rRNA (Martin *et al.*, 2013). For example, DDX5 functions in rRNA transcription and processing (Jalal, Uhlmann-Schiffler and Stahl, 2007; Saporita *et al.*, 2011), while DDX51 was shown to be necessary for normal cleavage of pre-rRNAs (Srivastava *et al.*, 2010). Consistent with these roles, we found that R-loop-interacting proteins DDX18, DDX24 and DDX27 localized largely to the nucleolus, with lower levels of localization to non-nucleolar chromatin (Fig. 2.3b). To further study the functions of several R-loop-interacting DEAD-box proteins, we tested the effects of DDX10, DDX24, DDX27, and DDX54 loss using RNA interference. Northern blotting revealed increased accumulation of different pre-rRNAs in different DEAD-box knockdown cells, suggesting these DEAD-box proteins impact rRNA maturation at several different steps. These data indicate that resolution of R-loops by multiple RNA helicases acting non-redundantly may be necessary to efficiently process the pre-rRNA transcript into mature rRNAs.

In addition, by performing mRNA-seq, we observed misregulation of a highly overlapping set of genes in all four KDs, including genes implicated in cellular differentiation and migration. These findings raise the possibility that DEAD-box helicases act in a common pathway to regulate the levels of hundreds

of mRNAs. These effects could potentially occur through direct effects on a shared set of mRNA encoding genes. Alternatively, the shared set of transcripts could be more sensitive to the change in rRNA levels, thus being misregulated upon knockdown of the four DEAD-box proteins.

Although stringent high salt extraction of chromatin prior to immunoprecipitation of R-loops may reduce contamination with general chromatin-binding proteins, this harsh treatment may also result in loss of dynamic or weakly-binding proteins that nevertheless play roles in R-loop regulation or function. To identify such proteins, we introduced a specific RNA-protein crosslinking step into our immunoprecipitation protocol and utilized SILAC labeling to increase sensitivity. By comparing the array of R-loop-interacting proteins identified with and without crosslinking, we observed that crosslinking increased enrichment of numerous proteins that were weakly enriched in the absence of crosslinking, as predicted, as well as a number of proteins that were not identified in the absence of crosslinking. For example, STRBP, a splicing factor known to bind to both DNA and RNA, was enriched by crosslinking. More interestingly, a large fraction of proteins highly enriched in the uncrosslinked dataset exhibited reduced enrichment in the presence of crosslinking, suggesting that many proteins that strongly interact with the RNA component of R-loops may become less easily detected by mass spectrometry approaches, due to the covalent addition of RNA nucleotides of unknown size. These two methods in combination enabled us to identify 364 stringent R-loop-interacting proteins, including known regulators of R-

loops and proteins previously not shown to bind these structures. Taken together, these findings suggest that both uncrosslinked and crosslinked S9.6 co-IP offer advantages for identification of the R-loop-interactome *in vivo*. These studies serve as a resource for uncovering the mechanisms by which R-loops are regulated, as well as the means by which R-loops might affect regulation or processing of RNA.

Materials and methods

Cell culture. E14 mESCs were maintained in tissue culture plates coated with 0.2% gelatin, in medium that contained DMEM-high glucose (MilliporeSigma, D6546-500ML), supplemented with 10% Fetal Bovine Serum (MilliporeSigma, F2442-500ML), 2 mM L-Glutamine (Corning, 25-005-CI), MEM Nonessential Amino Acids (Corning, 25-025-CI), β-mercaptoethanol (MilliporeSigma, M6250-500ML) and recombinant leukemia inhibitory factor.

Antibodies. Antibodies used included DDX18 (Bethyl Laboratories, A300-535A), DDX24 (Abcam, ab70463), DDX27 (Bethyl Laboratories, A302-216A), DDX54 (MilliporeSigma, AV36498-100UL), DHX9 (Abcam, ab26271), RPB1 (Santa Cruz Biotechnology, sc-899x), Fibrillarin (Novus Biologicals, NB300-269), CEBPZ (Proteintech, 25612-1-AP), SFPQ (Abcam, ab38148), CTCF (MilliporeSigma, 07-729), mouse IgG2a (Abcam, ab18413). The S9.6 monoclonal antibody was purified from the HB-8730 hybridoma, obtained from ATCC.

S9.6 co-immunoprecipitation. For uncrosslinked immunoprecipitation, mESCs were resuspended in 10 mM Tris-HCl pH 7.0, 150 mM NaCl, 0.15% NP-40 with 1 × Halt Protease Inhibitor Cocktail (Thermo Fisher Scientific, 78429), layered onto Sucrose Buffer (10 mM Tris-HCl pH 7.0, 150 mM NaCl, 25% sucrose, protease inhibitor) and centrifuged. The pellet was resuspended in stringent wash buffer (10 mM Tris-HCl pH 7.9, 1.5 mM $MgCl_2$, 420 mM NaCl, 25% glycerol, 0.2 mM EDTA, 0.5 mM DTT, protease inhibitor) and incubated on ice for 30 min. Nuclei were

centrifuged at 7000 g and resuspended in AM-150/0.1% NP-40 buffer (150 mM KCl, 20 mM Tris-HCl pH 7.9, 5 mM $MgCl_2$, 0.2 mM EDTA, 10% glycerol, 0.1% NP-40, with 1 × Halt Protease Inhibitor Cocktail). The salt-extracted nuclei were then sonicated using a Bioruptor (Diagenode) for 15 cycles, 30s on/30s off with the intensity set at medium. Based on protein concentration, 1 mg of sonicated chromatin was mixed with 20 µg S9.6 or mouse IgG2a as control and incubated overnight at 4 °C. 20 µg of sonicated chromatin was set aside untreated as Input (2% of IP samples). Before mixing sonicated chromatin with antibody, 0.2 µL RNase A (Thermo Fisher Scientific, EN0531) was added and treated at 37 °C for 15 min. For DNase I treatment, 10 µL DNase I (New England Biolabs, M0303L) was added and treated at 37 °C for 2 hr. The next day the mixture was incubated with pre-washed Protein G magnetic beads (New England Biolabs, S1430S), washed 3 times in AM-150/0.1% NP-40 buffer for 5 min each and eluted in 1 × SDS Loading Buffer (0.2 M Tris-HCl pH 6.8, 277 mM SDS, 40% glycerol, 6 mM bromophenol blue) by boiling for 10 min. For crosslinked IP, mESCs were cultured in SILAC medium (Thermo Fisher Scientific, A33972) either supplemented with standard lysine and arginine or $^{13}C_6$ $^{15}N_2$ lysine and $^{13}C_6$ $^{15}N_4$ arginine. 500 µM 4SU (Biosynth Carbosynth, NT06186) was added to heavy isotope treated cells for 2 hr while the light isotope cultured cells were treated with DMSO vehicle. Both sets of cells were then UV treated at a wavelength of 312 nm for 1J/cm^2, cells were lysed and combined after sonication at 1:1 ratio based on protein amount, and immunoprecipitated as above. After washing, IP samples were subjected to on-

beads RNase A treatment before elution as described above. Both uncrosslinked and crosslinked S9.6 co-IP were performed with three biological replicates.

LC-MS/MS. S9.6 co-IP elution was run on a 10% SDS-PAGE gel and gel slices were recovered with care to exclude the majority of the IgG heavy and light chains. Gel bands were in-gel digested and analyzed by LC-MS and LC-MS-MS as described previously (Chen, 2013; Chu *et al.*, 2019). The digestion mixture was separated on a 75 μm × 25 cm PepMap Rapid Separation Liquid Chromatography (RSLC) column (100 Å, 2 μm) at a flow rate of ~450 nL/min, and the eluant was analyzed by an LTQ Orbitrap XL mass spectrometer (Thermo Scientific, Waltham, MA). LC-MS data were acquired in a data-dependent acquisition mode, cycling between a MS scan (m/z 315-2,000) acquired in the Orbitrap, followed by collision-induced dissociation (CID) analysis on the 3 most intensely multiply charged precursors acquired in the linear ion trap. The centroided peak lists of the CID spectra were generated using PAVA searched against a database that is consisted of the Swiss-Prot protein database (version 2017.11.01, 16942/556006 entries searched for Mus Musculus), using Batch-Tag, a program of the University of California San Francisco Protein Prospector software, version 5.9.2. Protein hits were reported with a Protein Prospector protein score ≥ 22, a protein discriminant score ≥ 0.0 and a peptide expectation value ≤ 0.01 (Chalkley *et al.*, 2005). This set of thresholds of protein identification parameters did not return any substantial false positive protein hits from the randomized half of the concatenated database. Data are available via ProteomeXchange (Deutsch *et al.*, 2020) with identifier

PXD022697. For differential analysis of IP versus Input, we used the R package Prostar (Wieczorek *et al.*, 2017) to calculate the normalized fold change of spectra counting and p-value, and to make volcano plots.

For crosslinked co-IP, cells were differentially labeled, combined, immunoprecipitated, and processed for MS as described above, and fold change values were normalized based on the SILAC ratio. For comparison of the crosslinked and uncrosslinked S9.6 co-IP, we first included proteins that appeared in either or both of the input and immunoprecipitates in the uncrosslinked data (587 proteins), filtered this list to include only those proteins also found in the crosslinked dataset, and compared both datasets after Z-Score normalization.

Immunofluorescence staining. mESCs were cultured on gelatinized coverslips and fixed with 4% paraformaldehyde (Electron Microscopy Sciences, 15710) for 10 min. Cells were then permeabilized with 0.5% NP-40 and blocked with 5% Normal Goat Serum (Vector Laboratories, S-1000), 0.3% Triton X-100 (Amresco, M143-1L). Primary antibodies were diluted in 1% BSA, 0.3% Triton X-100 in PBS, the dilution were: DDX18 (1:250), DDX24 (1:50), DDX27 (1:250), Fibrillarin (1:250), SFPQ (1:250), CTCF (1:500). Secondary antibodies were Alexa Fluor 488 Goat anti-Rabbit IgG (Thermo Fisher Scientific, A-11008, 1:250) and Alexa Fluor 594 Goat anti-Mouse IgG (Thermo Fisher Scientific, A-11005, 1:250). DNA was stained by 1 μg/mL DAPI and slides were observed under Nikon Eclipse E400. IF were performed two to three independent times per antibody

Northern blotting. DNA probes were amplified using cDNA from mESCs as a template, radiolabeled by [α-32P]dCTP (PerkinElmer, 3000Ci/mmol 10mCi/mL, BLU013H100UC) using Prime-a-Gene Labeling System (Promega, U1100). Radioactivity was determined by a scintillation counter, 5 X 10^6 cpm of labeled probes were used for a Northern blotting assay. 1 μg of total RNA was load into gel containing 1% agarose and 6% formaldehyde, ran in 1 X MOPS buffer 100 volts for 1.5 hr. RNA was transferred from gel to Amersham Hybond-N+ membrane using 10 X SSC buffer (Invitrogen, 15557044) overnight. The next day, blots were crosslinked at a wavelength of 254 nm for 120 mJ/cm^2. Blots were prehybridized with 10 ml PerfectHyb Plus hybridization buffer (MilliporeSigma, H7033-50ML) at 68 °C for 10 min. Radiolabeled probes were added, hybridized at 68 °C overnight with rotation. The next day, blots were quickly washed twice with 5 mL 2x SSC-0.1% SDS Wash Buffer, then washed with 10 mL 2x SSC-0.1% SDS Wash Buffer for 10 min. The blots were exposed to X-ray film. Two biological replicates were performed with similar results.

Western blotting. Dilution for antibodies was: DDX18 (1:4,000), DDX27 (1:1,000), DDX54 (1:500), DHX9 (1:1,000), RPB1 (1:1,000), CEBPZ (1:2,000), SFPQ (1:1,000). Western Blotting was quantified by ChemiDoc Touch Imaging System (Bio-Rad).

EsiRNA preparation and transfection. Endoribonuclease-prepared small interfering RNA (esiRNAs) were prepared as described(Fazzio, Huff and Panning, 2008). Briefly, cDNA from mESCs was used as template to amplify cDNA that

targets each gene with T7 anchor sequence added, *in vitro* transcription was then performed using T7 RNA Polymerase (New England Biolabs, M0251L). RNA was digested by ShortCut RNase III (New England Biolabs, M0245L) to generate a pool of small siRNAs, which was purified using a PureLink RNA Mini Kit (Thermo Fisher Scientific, 12183020). For transfection, 400 ng esiRNA was mixed with 0.4 mL serum-free medium and 4 μL Lipofectamine 2000 (Thermo Fisher Scientific, 11668-019), and after 15 min of incubation 2.8×10^5 mESCs were added and the mixture was plated in one well of a gelatinized 6-well plate. Media was replaced ~16 hr later and cells were harvested 48 hr after transfection.

RT-qPCR and RNA-seq. RNA was extracted using an RNA Clean & Concentrator-25 Kit (Zymo Research, 11-353B) with on-column DNase I digestion for 1 hr. For RT-qPCR, cDNA was synthesized using purified MMTV reverse transcriptase. Quantification was performed using primers targeting cDNA, with *Gapdh* used as a loading control (Table 2.4). We performed three biological replicates each with three technical replicates. RNA-seq libraries were prepared by BGI Genomics Company. mRNA was enriched from total RNA by oligo(dT)-attached magnetic beads, followed by fragmentation, first- and second-strand cDNA synthesis, end repair, add A, adapter ligation, and PCR amplification. PCR products were purified with Ampure XP beads (Beckman Coulter, A63881). The PCR products were sequenced on BGISEQ-500 using 100-base paired-end sequencing. Reference genome mapping, transcript assignment, quantification, and differential analysis were done using RSEM (Li and Dewey, 2011). Heatmaps were made by Java

TreeView. GO term analyses were done by PANTHER Classification System, and overlapping of RNA-seq data was illustrated by UpSet (Lex *et al.*, 2014). Three biological replicates were performed for RNA-sequencing.

Data availability. Primary mass spectrometry data are available via ProteomeXchange with identifier PXD022697. RNA-sequencing data are deposited in Gene Expression Omnibus (GSE161890).

Table 2.4. Primers for preparing esiRNAs and performing RT-qPCR

Primer sets	Sequence
Ddx10_esiRNA_F	GGGCGGGTTGGCCACAGATTCAGAAATG
Ddx10_esiRNA_R	GGGCGGGTCCAGCTCTTCATCCTCTGCT
Ddx24_esiRNA_F	GGGCGGGTAGTGGAGACGCTAACGGAGA
Ddx24_esiRNA_R	GGGCGGGTCTTGGACTGCACTGGAAACA
Ddx27_esiRNA_F	GGGCGGGTAGCAGAAAGCCTTGCAGAAG
Ddx27_esiRNA_R	GGGCGGGTGCTGTAATGGCCTTCAGGAG
Ddx54_esiRNA_F	GGGCGGGTTGTTCGATGAAGCAGACAGG
Ddx54_esiRNA_R	GGGCGGGTAGTCACGATGAGGGTGGAAC
Gapdh_RTqPCR_F	TTGATGGCAACAATCTCCAC
Gapdh_RTqPCR_R	CGTCCCGTAGACAAAATGGT
Cdkn1a_RTqPCR_F	GTGGCCTTGTCGCTGTCTTG
Cdkn1a_RTqPCR_R	CAATCTGCGCTTGGAGTGATAGAAA
Cdkn2b_RTqPCR_F	GCAGATCCCAACGCCCTGAA
Cdkn2b_RTqPCR_R	GGTCTGGTAAGGGTGGCAGG
Wnt5a_RTqPCR_F	TGCCATGTCTTCCAAGTTCTTCC
Wnt5a_RTqPCR_R	GGGTTATTCATACCTAGAGACCACCA
Pcsk9_RTqPCR_F	AAGATGAGCAGTGACCTGTTGGG
Pcsk9_RTqPCR_R	GCGGTCTTCCTCTGTCTGGTG

CHAPTER III: Characterization of R-loop-interacting protein CEBPZ reveals roles in R-loop regulation and colocalization with CTCF

Preface

Data presented in this chapter are unpublished work performed by Tong Wu and Thomas Fazzio.

Author contributions. Tong Wu and Thomas Fazzio designed experiments. Tong Wu performed most of the experiments with help from Thomas Fazzio.

Abstract

CEBPZ is a transcription factor that was recently found to control the abundance of m6A, an mRNA modification that has been shown to promote R-loop formation. I identified CEBPZ as an R-loop-associated protein by R-loop immunoprecipitation followed by mass spectrometry. To test for potential roles of CEBPZ in regulation of R-loops, I first performed CUT&RUN to identify its genomic localization. By comparing with DRIP-seq that showed R-loop distribution, some R-loops were shown to localize close to CEBPZ binding sites, which were found down-regulated by CEBPZ depletion. Moreover, unexpected colocalization of CEBPZ with CTCF, a protein known for regulating chromatin looping, was revealed by comparing their localization in the genome, with their interaction further confirmed by co-IP. However, conflicting results from CUT&RUN and ChIP-seq were observed when investigating the effect of CEBPZ depletion on CTCF binding. H3K4me3 CUT&RUN confirmed that CEBPZ depletion affected CUT&RUN performance but not CTCF and H3K4me3 binding in the genome. This discrepancy raises interesting possibilities about how CEBPZ loss has non-specific effects on CUT&RUN mapping and what these signify. These possibilities are still under study.

Introduction

R-loops are co-transcriptional nucleic acid structures, with most well-studied R-loops dependent on RNAP II transcription (Ginno *et al.*, 2012; Sanz *et al.*, 2016). Despite the outsized emphasis on RNAP II-dependent R-loops, R-loops formed during rDNA transcription are a major source of them in cells. rDNA contains 200-600 copies in humans and 70-400 copies in mice (Parks *et al.*, 2018). rDNA and its transcribed rRNA (28S, 18S, 5.8S) form nucleolus. These rRNAs contribute to about 80% of total RNA in cells (Harvey *et al.*, 2000), thus forming a high-abundant of R-loops observed in the nucleus. In yeast, R-loops accumulate across 18S and 28S rDNA, especially enrich at the 5' end of 18S and 28S (El Hage *et al.*, 2010). Their accumulation dramatically increases when Top1 and Top2, topoisomerases that resolve negative supercoiling, are depleted (El Hage *et al.*, 2010). In HeLa cells, IF of TOP1 and ribonuclease RNASEH1 shows their strong localization to nucleoli and colocalization with S9.6 signal. Inhibition of RNAP I transcription abolishes the S9.6 IF signal observed in nucleoli, and also causes migration of TOP1 and RNASEH1 to the perinucleolar region, indicating their strong binding affinity for the high abundance of R-loops in nucleoli (Shen *et al.*, 2017). Moreover, accumulation of R-loops has also been linked to change or disruption of nucleolar structure (Shen *et al.*, 2017; Zhou *et al.*, 2020). Increasing R-loop levels by knocking down *RNASEH1* causes fragmented nucleoli. Inhibition of rRNA transcription elongation causes enhanced rRNA transcription initiation

and accumulation of R-loops at the 5’ end of rDNA, leading to fragmented nucleoli (Zhou *et al.*, 2020).

CEBPZ was first identified in a screen for factors that bind to the CCAAT element of the *hsp70* promoter. It has sequence-specific binding activity to the CCAAT sequence as well as transcription stimulation activity of the *hsp70* gene (Lum *et al.*, 1990). Later, Imbriano *et al.* showed that the CEBPZ binding to CCAAT boxes from the *hsp70* and *hsp40* promoters requires another transcription factor named NF-Y. Their research indicates that NF-Y directly interacts with CCAAT sequences and recruits CEBPZ to its binding sites, and CEBPZ does not have the ability to activate *hsp70* or *hsp40* genes without the presence of NF-Y (Imbriano *et al.*, 2001).

CEBPZ has also been shown to bind at the same genomic regions with METTL3. METTL3 is the catalytic subunit of the N6-methyltransferase complex that methylates adenosine of mRNA to form N6-methyladenosine (m6A), which is the most abundant co-transcriptional mRNA modification identified in eukaryotes. *CEBPZ* KD induces decrease of METTL3 at regions where CEBPZ and METTL3 bind. Moreover, *CEBPZ* KD has also been shown to cause reduction of m6A level (Barbieri *et al.*, 2017). m6A modification has been shown to function in various processes, including alternative splicing, RNA export, translation, and RNA degradation (Zhang, Fu and Zhou, 2019). In recent years, functional roles of m6A in the regulation of R-loops have been described. Using liquid chromatography with tandem mass spectrometry, Abakir *et al.* show that m6A but not other RNA

modifications are associated with RNA of RDHs in human induced pluripotent stem cells. m6A DNA immunoprecipitation, which identifies m6A distribution in RNA that is associated with the genome, shows presence of m6A in the majority of RDHs (Abakir *et al.*, 2019). Mutation of adenosine to uridine in a minigene reporter system causes reduction of m6A and R-loop level and facilitates RNAP II readthrough. (Yang *et al.*, 2019). In all, m6A has been shown to be present in RNA of R-loops and promote R-loop formation.

CEBPZ was one of the transcription factors identified in both our uncrosslinked and crosslinked MS studies. Considering the role of CEBPZ in the regulation of m6A modification and the presence of m6A in R-loops, in this chapter, I examine the genomic distribution of CEBPZ, and whether CEBPZ regulates levels of R-loops associated with mRNA and rRNA.

Results

CEBPZ colocalizes with R-loops and regulates their levels

To further study the R-loop-interacting proteins in the regulation of R-loops, I chose one of the transcription factors enriched by S9.6 named CEBPZ for closer examination. Its enrichment by S9.6 co-IP was confirmed by Western blotting (Fig. 3.1a). CEBPZ recognizes the CCAAT binding motif present at many promoters to drive gene expression. In addition, CEBPZ has been shown to regulate the recruitment of METTL3 to the genome (Barbieri *et al.*, 2017). METTL3 methylates adenosine on RNA to form m6A in a co-transcriptional manner. Importantly, METTL3 and m6A have been shown to promote R-loop formation. Therefore, I hypothesized that depletion of CEBPZ would cause reduction of R-loops with which it associates.

To identify the R-loops that could potentially be affected by CEBPZ depletion, I first tried to identify the genomic distribution of CEBPZ. I took advantage of the cleavage under targets and release using nuclease (CUT&RUN) method to study the CEBPZ binding in the genome (Fig. 3.1b). Serving as an alternative method for traditional ChIP-seq for mapping factors that bind genome, CUT&RUN does not require crosslinking, produces less background, and can easily be adapted to low cell numbers (Skene and Henikoff, 2017). CEBPZ CUT&RUN produced ~2000 peaks relative to no antibody control libraries. To illustrate one example, at the intron region of the *Gphn* gene, CEBPZ showed a

peak over no antibody control (no ab), which overlapped with an R-loop peak revealed by DRIP-RNA-seq (DRIP-seq like technique with the RNA part of R-loops being sequenced) performed in mESCs from a previous study (Fig. 3.1c) (Chen *et al.*, 2015). By overlapping peaks identified in the two datasets, 120 peaks from CUT&RUN and DRIP were found to overlap or be present within 1 kb of each other (Fig. 3.1d).

Next, I wanted to study the effect of CEBPZ depletion on the formation of R-loops. To alter the levels of CEBPZ in cells, I took advantage of the auxin-inducible degron (AID) system (Nishimura *et al.*, 2009). We obtained an mESC line engineered to express the TIR1 F-box protein from *Oryza sativa* (rice), and tagged CEBPZ at its endogenous locus with an auxin-inducible destabilizing degron and a hexahistidine triple FLAG tag at the C-terminus. TIR1 can form a functional ubiquitin E3 ligase composed in complex with other proteins from mESCs. Upon addition of an auxin derivative, indole-3-acetic acid (IAA), the TIR1 ubiquitin ligase ubiquitylated the AID tagged CEBPZ, resulting in its degradation (Fig. 3.1e). Western blotting using CEBPZ antibody showed that by incubating IAA with mESCs for 6 hours, a partial decrease of AID tagged CEBPZ can be observed. 24 and 48 hours of incubation brought CEBPZ to a very low level, without affecting the expression of any other proteins examined, such as β-actin (Fig. 3.1f).

I then performed DRIP-qPCR with or without IAA treatment for 48 hours to study the CEBPZ regulation of R-loop abundance. By performing DRIP-qPCR using primers that target several R-loops that do not exhibit CEBPZ binding within

5 kb (*Wipf2*, *Sp2*, *Sp1*), or R-loops that have CEBPZ binding within 1 kb distance (*Jarid2*, *Gphn*), I was able to show that upon CEBPZ depletion, CEBPZ-proximal R-loops showed a more dramatic reduction in R-loop levels relative to R-loops far from any CEBPZ binding site (Fig. 3.1g). This indicated that CEBPZ is a potential R-loop regulator, which is consistent with my hypothesis.

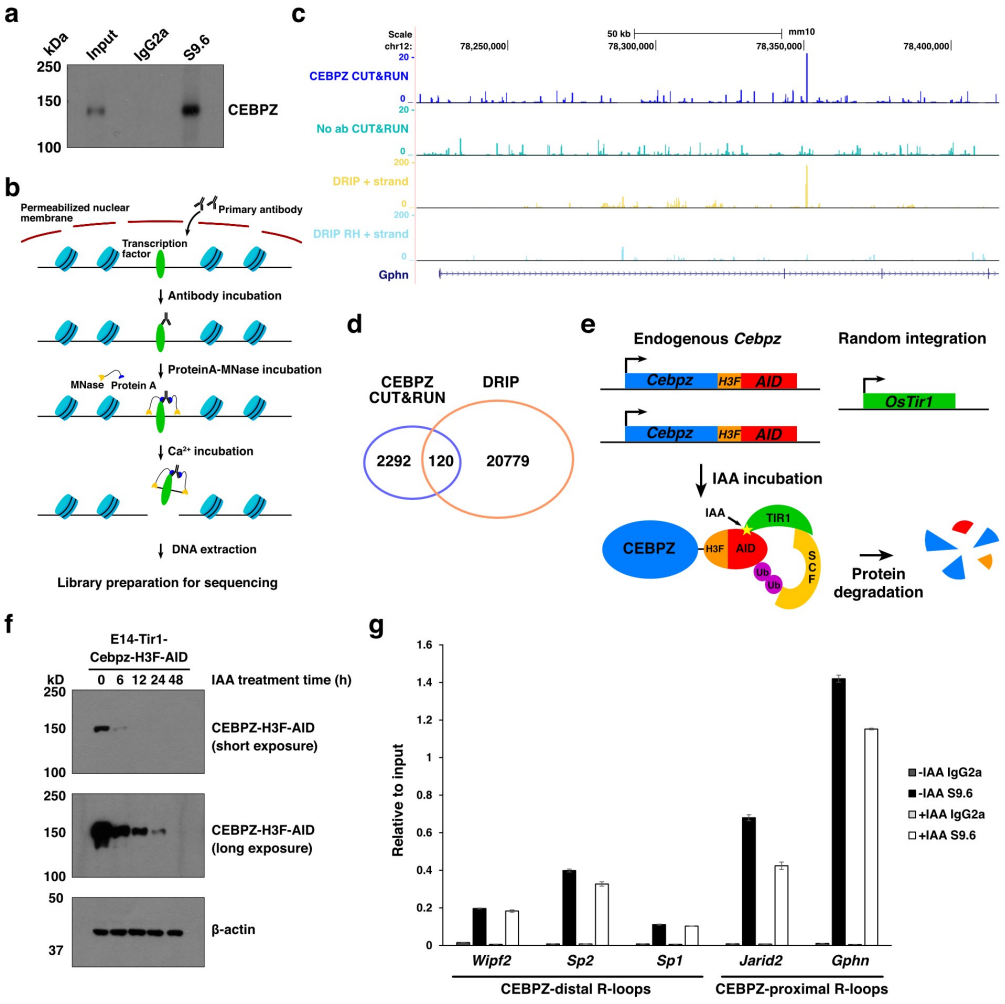

Figure 3.1 CEBPZ colocalizes with R-loops and regulates them.

a. Western blot of CEBPZ protein enriched by S9.6 co-IP. **b.** Schematic diagram of CUT&RUN experiment. **c.** Browser track comparing CUT&RUN and DRIP-RNA-seq results on *Gphn* gene. No antibody (light blue) was used as a control for CEBPZ CUT&RUN (blue). RNase H (RH) treated DRIP-RNA-seq (faint blue) was used as a control for untreated DRIP-RNA-seq (yellow), with only plus-strand

displaced. **d.** Venn diagram of the overlap between the peaks identified by CEBPZ CUT&RUN and DRIP-RNA-seq, peaks from two experiments that their center is within 1 kb of each other were counted as shared peaks. **e.** Schematic diagram of the AID system for depletion of CEBPZ. **f.** Depletion of CEBPZ tagged by hexahistidine triple Flag tag and AID tag. Upon IAA treatment for 0, 6, 12, 24, and 48 hours, CEBPZ was detected by CEBPZ antibody, with β-actin as a loading control. **g.** DRIP-qPCR performed upon CEBPZ depletion. Five genomic regions that contain R-loops were targeted.

Colocalization of CEBPZ and transcription factor CTCF in the genome

To gain additional insights into the functions of CEBPZ, I next performed motif enrichment analysis of regions enriched for CEBPZ binding by CUT&RUN. The most significantly enriched sequence was the CCAAT binding motif corresponding to NF-Y, the binding partner of CEBPZ, confirming the specificity of the CUT&RUN data (Fig. 3.2a). Surprisingly, the binding motif that belongs to BORIS/CTCFL was also very highly ranked (Fig. 3.2a). Brother of the regulator of the imprinted site (BORIS), also known as CCCTC-binding factor like (CTCFL), is a paralog of the insulator binding protein CTCF. BORIS/CTCFL and CTCF share highly similar 11-zinc-finger DNA binding domains as well as similar DNA binding motifs (Loukinov *et al.*, 2002).

CTCF is well-known for its roles in 3D genome organization, as well as gene "insulation": segregation of enhancer regulatory units away from nearby genes to prevent promiscuous enhancer utilization. Even though different groups have examined CTCF functions in multiple organisms, CEBPZ was not reported as a CTCF interactor, nor has it been shown to bind at the same genomic loci as CTCF (Marino *et al.*, 2019; Lehman *et al.*, 2021). To study if CEBPZ and CTCF share the same binding regions in the genome, I performed motif analysis by CTCF CUT&RUN. CTCF binding sites were enriched for the CTCF binding motif, as well as the CCAAT motif shared by NF-Y and CEBPZ (Fig. 3.2b). A browser track of CEBPZ and CTCF CUT&RUN shows their frequent colocalization in the genome, for example, at the promoter of the *G3bp1* gene (Fig. 3.2c). Indeed, of the 1942

CEBPZ peaks identified by CUT&RUN, 1557 have CTCF binding within 1 kb distance, while most CTCF binding sites do not exhibit CEBPZ binding (Fig. 3.2d). This suggested that most CEBPZ binding sites are also CTCF sites, but not the reverse.

To further study the presence of CEBPZ and CTCF at some shared genomic target loci, I next investigated whether CEBPZ and CTCF physically interact. By performing IP of CEBPZ, CTCF was shown to be pulled down by CEBPZ antibody in WT mESCs, indicating their potential interactions (Fig. 3.2e). The reverse IP using CTCF antibody to IP CEBPZ was not able to identify detectable levels of CEBPZ by Western blotting (data not shown). This may be due to that the majority of CEBPZ interacts with CTCF, but only a small portion of CTCF has CEBPZ binds, which was indicated by the CUT&RUN data (Fig. 3.2d). Moreover, by using an antibody that recognizes hexahistidine tag of CEBPZ in the Cebpz-H3F-AID cell line, CTCF was precipitated together with CEBPZ-H3F-AID, though at a different IP efficiency than observed in WT cells (Fig. 3.2e).

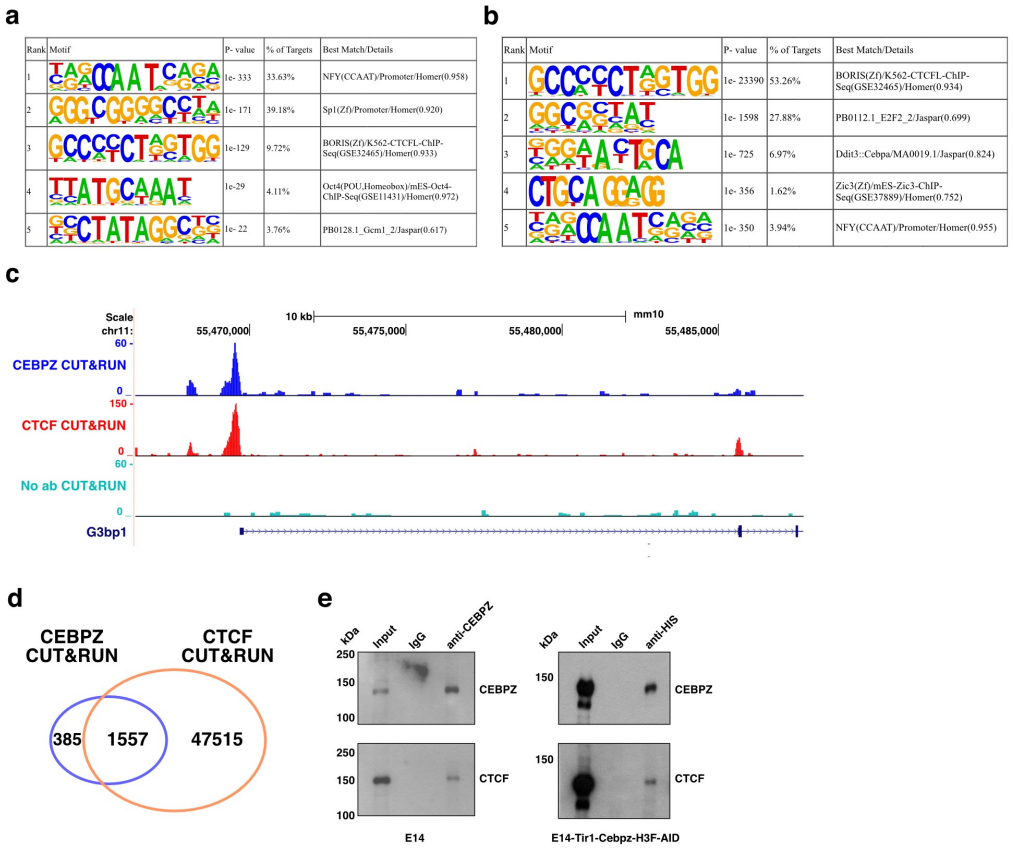

Figure 3.2. CEBPZ and CTCF bind to some shared genomic regions.

a. HOMER *de novo* motif analysis of CEBPZ CUT&RUN with top five motifs. Motif sequence, percent of targets, p-value, and motif details were shown. **b.** HOMER *de novo* motif analysis of CTCF CUT&RUN, as depicted in a. **c.** Browser tracks of CEBPZ (blue) and CTCF CUT&RUN (red), with no antibody as a control (light blue). The promoter of *G3bp1* was used as a representative region. **d.** Venn diagram of the overlap between the peaks identified by CEBPZ CUT&RUN and CTCF CUT&RUN, peaks from two experiments that their center is within 1 kb of

each other were counted as shared peaks. **e.** Western blotting of CEBPZ and CTCF immunoprecipitated by CEBPZ antibody in the WT mESCs or anti-His antibody in the Cebpz-H3F-AID cell line.

Effect of CEBPZ depletion on the performance of CUT&RUN

Since CEBPZ interacts with CTCF and colocalizes at shared genomic regions, I hypothesized that CEBPZ might recruit CTCF to some genomic regions. In this scenario, depletion of CEBPZ would likely cause a reduction in CTCF binding, particularly at CEBPZ binding sites. By treating the Cebpz-H3F-AID line with IAA for 24 hours, Western blotting showed depletion of CEBPZ with CTCF level unchanged (Fig. 3.3a). CTCF CUT&RUN showed that average CTCF binding over all of its binding sites dramatically decreased upon depletion of CEBPZ (Fig. 3.3b). This reduction in CTCF binding occurred not only at the CTCF binding sites that overlap with CEBPZ, but also at CTCF binding sites that lack CEBPZ binding. These results suggested that either CEBPZ had a global impact on chromatin association of CTCF, or CEBPZ depletion reduced the efficiency of CUT&RUN.

To distinguish between these two possibilities, I next performed CTCF ChIP-seq, a well-established technique for mapping binding sites of proteins. Compared to CUT&RUN, ChIP-seq requires crosslinking of the proteins to genomic DNA and physical fragmentation of genome using sonication (Baranello *et al.*, 2016). In contrast, CUT&RUN detects proteins binding sites in a more native environment, which may be affected by global chromatin architecture. Interestingly, CTCF ChIP-seq showed that upon CEBPZ depletion, there is essentially no reduction of CTCF binding at its binding sites, in conflict with what I observed by CUT&RUN (Fig. 3.3c). To investigate this discrepancy, I performed CUT&RUN of H3K4me3, a histone modification enriched at active promoters

(Howe *et al.*, 2017). CEBPZ has not been reported to impact H3K4me3 deposition. Similar to CTCF, H3K4me3 CUT&RUN showed a dramatic decrease of H3K4me3 occupancy near TSSs (Fig. 3.3d). This suggested that CEBPZ depletion may affect the performance of CUT&RUN such that CTCF and H3K4me3 CUT&RUN experiments recover relatively fewer reads mapped to their binding sites upon CEBPZ depletion. Interestingly, we found that upon CEBPZ depletion, larger quantities of DNA are recovered after library preparation, suggesting CEBPZ depletion causes elevated MNase digestion at genomic regions not bound by CTCF or H3K4me3. Consequently, enrichment at CTCF or H3K4me3 sites appeared to be lower upon CEBPZ depletion due simply to an increase in the background in these libraries.

The CUT&RUN technique identifies protein binding in the genome in a more native status compared to ChIP-seq, which means it may be affected by the biophysical features of chromatin (e.g., fluidity) or chromatin architecture (e.g., chromatin “openness”). To see if CEBPZ depletion affects chromatin architecture, which in turn may affect CUT&RUN performance, I performed the assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq), a technique for mapping chromatin accessibility (Buenrostro *et al.*, 2015). Upon CEBPZ depletion, ATAC-seq revealed a slight decrease in chromatin accessibility near CTCF binding sites (Fig. 3.3e), although this small difference may be insufficient to explain the CUT&RUN data. Another possibility is that CEBPZ depletion changes three-dimensional chromatin architecture and/or nuclear

structure, which in turn affects CUT&RUN performance. However, the distribution and pattern of CTCF (Fig. 3.3f) and H3K4me3 (Fig. 3.3g) staining measured by IF showed no observable change upon IAA treatment. Similarly, genomic DNA stained by DAPI showed minimal differences in nuclear morphology in CEBPZ depleted cells. This confirmed that CEBPZ depletion did not cause a dramatic change of chromatin or nuclear structure, nor did it cause relocalization of the proteins tested. At this stage, additional assays must be carried out to address the mechanisms by which CEBPZ affects CUT&RUN performance and what the significance of this finding might be. Examples of such approaches could be investigating changes in DNA lesions (e.g., breaks or nicks) or alterations in nucleosome positioning in CEBPZ depleted cells.

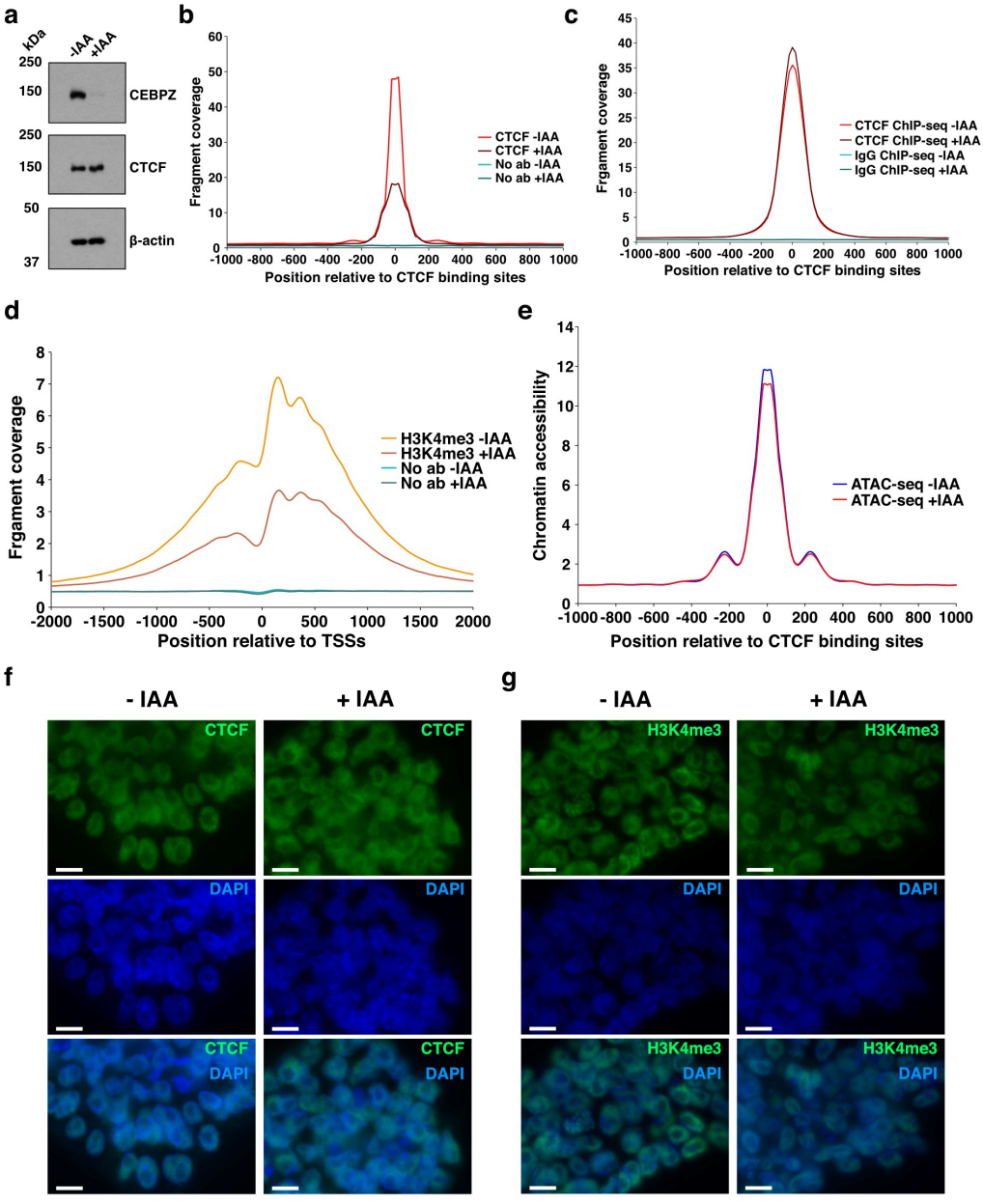

Figure 3.3. CEBPZ depletion affects CUT&RUN performance.

a. Western blotting of CEBPZ and CTCF upon 24 hours of IAA treatment, with β-actin as a loading control. **b.** Aggregation plot of CTCF CUT&RUN signal mapped

around CTCF binding sites upon CEBPZ depletion, no antibody CUT&RUN as a control, ± 1 kb of flanking regions were shown. **c.** Aggregation plot of CTCF ChIP-seq mapped around CTCF binding sites upon CEBPZ depletion, rabbit IgG was used as a control. **d.** Aggregation plot of H3K4me3 CUT&RUN mapped around TSSs upon CEBPZ depletion, ± 2 kb of flanking regions were shown. **e.** Aggregation plot of ATAC-seq signal mapped around CTCF binding sites upon CEBPZ depletion. **f.** Immunofluorescence staining of CTCF with or without IAA treatment. DNA was stained by DAPI. Scale bar = 10 μm. **g.** Immunofluorescence staining of H3K4me3 with or without IAA treatment, as depicted in f.

Regulation of rRNA abundance and rRNA-associated R-loops by CEBPZ

When performing IF using the anti-Flag antibody that targets the epitope tag at the C-terminus of CEBPZ, a nucleolar staining pattern was observed, as indicated by its colocalization with nucleolar marker Fibrillarin (FBRL) (Fig. 3.4a). The nucleolar IF signal was specific for CEBPZ, as indicated by IAA treatment in the Cebpz-H3F-AID line, which dramatically diminished the staining (Fig. 3.4b).

Based on the IF of CEBPZ, I then examined the regulatory roles of CEBPZ in rRNA transcription and processing. Using primers that targeted the 18S and 28S rRNA sequences, RT-qPCR showed that the level of the 18S- and 28S-containing rRNAs gradually decreased upon 12, 24, 48 hours of IAA treatment. In contrast, RT-qPCR using primers that target the 5' external transcribed spacer (5' ETS), which only exists in the nascent 47S pre-rRNA, revealed no change upon IAA treatment (Fig. 3.4c). These data suggested that CEBPZ plays roles in rRNA processing but not transcription. Moreover, DRIP-qPCR performed using primers that target the 18S and 28S rDNA regions showed a decrease in R-loop levels at those regions upon depletion of CEBPZ (Fig. 3.4d). This decrease may be caused by a defect in rRNA processing upon CEBPZ depletion that leads to lower levels of pre-rRNA associated with rDNA.

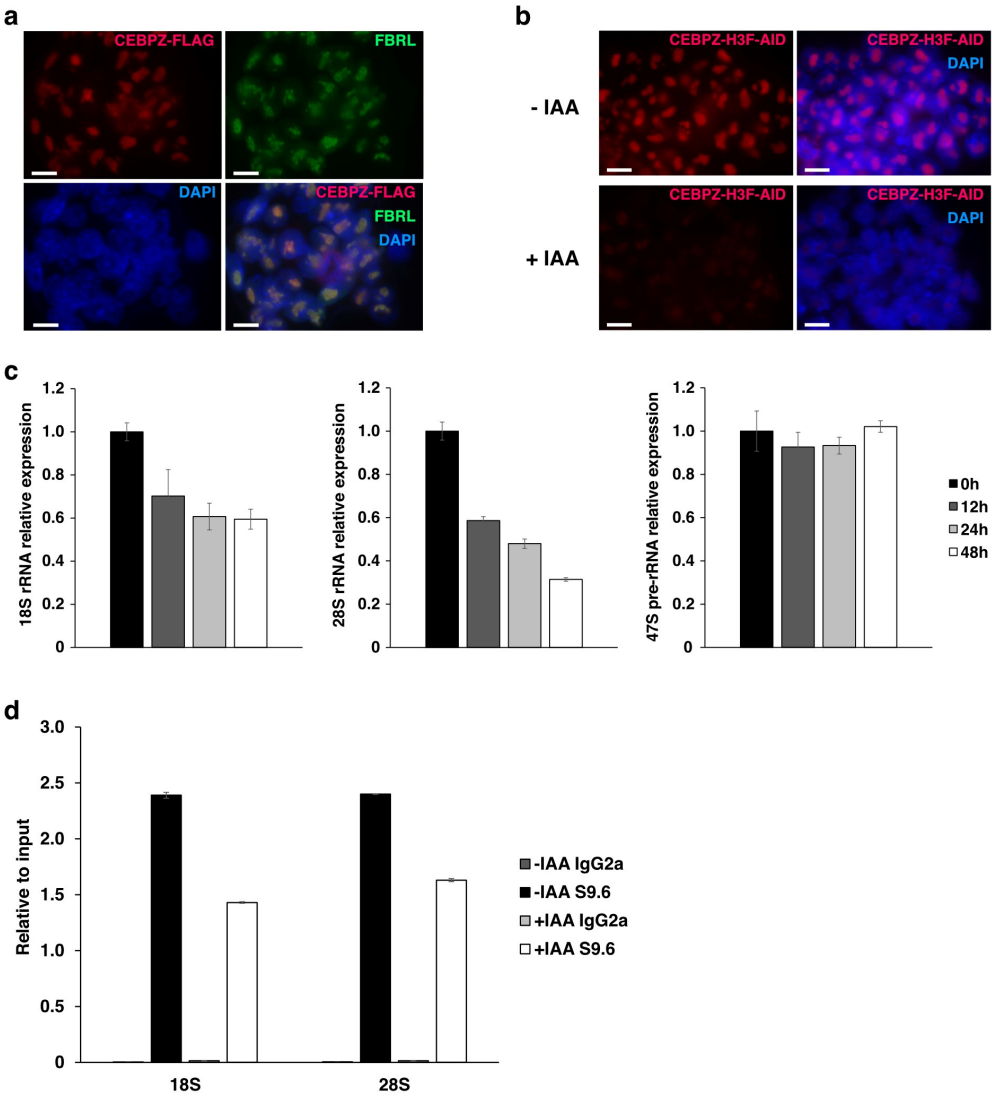

Figure 3.4. CEBPZ regulates rRNA processing and rRNA-associated R-loops.

a. Immunofluorescence staining of CEBPZ-3×Flag by Flag tag antibody, co-stained with FBRL antibody to mark the nucleolus. DNA was stained by DAPI. Scale bar = 10 μm. **b.** Immunofluorescence staining of CEBPZ-6×His-3×Flag-AID by Flag tag antibody, with or without IAA treatment. DNA was stained by DAPI.

Scale bar = 10 μm. **c.** Relative levels of rRNA, after IAA treatment of 0, 12, 24, or 48 hours. Levels in 0 hour were set to 1 after normalization to *Gapdh*. Three technical replicates were included. From left to right: rRNA containing 18S sequence, rRNA containing 28S sequence, pre-rRNA containing 5' ETS sequence. **d.** DRIP-qPCR performed upon CEBPZ depletion. R-loops at 18S and 28S rDNA regions were targeted.

Discussion

CEBPZ is a transcription factor we found to be enriched on or near R-loops, using an IP/mass spectrometry approach. To study the regulation of R-loops by CEBPZ, I performed CEBPZ CUT&RUN to identify its genomic distribution relative to R-loops. CUT&RUN revealed partial colocalization of CEBPZ peaks with R-loop peaks in the genome. By introducing an AID tag at the C-terminus of *Cebpz*, depletion of CEBPZ was achieved upon IAA treatment. DRIP-qPCR showed that R-loop levels decreased upon depletion of CEBPZ, especially at genomic regions where there is overlapping of CEBPZ binding with R-loops. This agrees with the established function of CEBPZ in the regulation of m6A on mRNA through recruitment m6A writer METTL3 (Barbieri *et al.*, 2017), and the function of m6A to promote R-loop formation (Abakir *et al.*, 2019; Yang *et al.*, 2019). To further solidify the role of CEBPZ in R-loop abundance, METTL3 recruitment to CEBPZ-binding genes and m6A levels on transcripts of genes bound by CEBPZ will need to be investigated upon CEBPZ depletion.

Motif analysis of CEBPZ CUT&RUN showed its CCAAT binding motif. Surprisingly, the binding motif of CTCF was within the top five motifs being identified. CTCF is a transcription factor mostly known for regulating 3D genome folding by forming boundaries for loops and TADs. R-loops have previously been shown to regulate CTCF binding at certain TAD boundaries (Luo *et al.*, 2020), and CTCF was enriched by S9.6 co-IP in this thesis. However, no clear colocalization or physical interaction of CTCF and CEBPZ in the genome has been reported. To

further confirm their colocalization in the genome, CTCF CUT&RUN was performed, motif analysis revealed its binding motif and CEBPZ binding motif. By overlapping CEBPZ with CTCF CUT&RUN, a high number of CEBPZ peaks were found to be near CTCF binding sites, but not the reverse. In agreement with these findings, IP of CEBPZ resulted in co-immunoprecipitation of CTCF, but not the reverse. This may suggest that a high portion of CEBPZ binds with CTCF, but only a small portion of CTCF binds CEBPZ.

To further study if CEBPZ recruits CTCF to regions where they colocalize, I performed CTCF CUT&RUN upon CEBPZ depletion. It showed that CTCF binding in the genome significantly decreased upon CEBPZ depletion. However, this conclusion was not supported by CTCF ChIP-seq, which revealed minimal changes in CTCF enrichment upon CEBPZ depletion. To better understand the discrepancy between CUT&RUN and ChIP-seq, H3K4me3 distribution in the genome was measured by CUT&RUN, which also showed a significant decrease upon CEBPZ depletion. H3K4me3 has not been shown by any group to be regulated by CEBPZ, supporting the idea that CEBPZ depletion affects CUT&RUN performance in general, rather than CTCF binding and H3K4me3 enrichment specifically. Together with the observation that CEBPZ depletion caused me to recover higher levels of DNA in CUT&RUN libraries, these results confirmed that CTCF and H3K4me3 occupancy in the genome were not decreased upon CEBPZ depletion. It raised the possibility that CEBPZ depletion seemed to increase the

levels of background DNA fragments released by MNase in the CUT&RUN procedure.

To investigate the possible mechanisms by which CEBPZ would affect CUT&RUN performance, I performed ATAC-seq to study chromatin openness. ATAC-seq signal showed a slight decrease of accessibility when CEBPZ was depleted. Reduction of chromatin accessibility would lead to a slight reduction in DNA recovery in CUT&RUN libraries rather than an increase. Therefore, these data are inconsistent with the possibility that changes in chromatin accessibility in CEBPZ depleted cells can explain the CUT&RUN results. Moreover, IF of CTCF and H3K4me3, along with DAPI staining of nuclei, showed minimal changes upon CEBPZ depletion, arguing against the possibility that chromatin structure or nuclear structure dramatically changed by CEBPZ depletion to cause mis-performance of CUT&RUN. At this time, I am still looking for possible mechanisms that explain this phenomenon, for example, by examining DNA damage markers as more DNA damage may cause more DNA fragments to be released upon MNase digestion. I also plan to examine nucleosome occupancy in the CEBPZ depleted cells, as the distribution of nucleosomes in the genome may affect pA-MNase binding to primary antibodies and fragmentation of genomic DNA.

IF of CEBPZ showed strong nucleolar localization and weaker nuclear signal, suggesting it may function primarily at the rDNA, despite its established role as a transcription factor. RT-qPCR showed that depletion of CEBPZ affected processing but not transcription of rRNAs. Consistent with this finding, DRIP-qPCR

showed that CEBPZ depletion caused downregulation of R-loops at rDNA regions, possibly through its effects on rRNA abundance. This research sheds light on a new role of CEBPZ within the nucleolus as a regulator of rRNA, and its function in regulating R-loops associated with rRNA and mRNA transcription. Further research is required to examine whether or not CEBPZ regulates rRNA and mRNA through similar mechanisms, and whether m6A modification is critical for the function of CEBPZ at both types of transcripts. This will increase our understanding of the roles of CEBPZ in the regulation of R-loops and the importance of this regulation within cells.

Materials and methods

Antibodies. Antibodies used included: Fibrillarin (Novus Biologicals, NB300-269), CEBPZ (Proteintech, 25612-1-AP), CTCF (MilliporeSigma, 07-729), H3K4me3 (MilliporeSigma, 05-745R), Flag Tag (MilliporeSigma, F1804), Histidine Tag (MilliporeSigma, 05-949), mouse IgG2a (Abcam, ab18413), rabbit IgG (Abcam, ab37415). The S9.6 monoclonal antibody was purified from the HB-8730 hybridoma, obtained from ATCC.

S9.6 co-immunoprecipitation. Uncrosslinked immunoprecipitation was performed as described in Chapter II. S9.6 co-IP followed by Western blotting was performed with two biological replicates.

CRISPR-Cas9 for genome editing. gRNA (Table 3.1) targeting the C-terminus of the *Cebpz* gene was cloned to the vector containing a spCas9 expression cassette and a puromycin-resistant marker. gBlock dsDNA (Table 3.2) that contains the 6×His-3×Flag-AID sequence with *Cebpz* C-terminus homology arms on both sides were cloned to a vector. 3 μg of both plasmids were mixed with 100 μL OptiMEM (Gibco, 31985070), 24 μL Fugene HD (Promega, E2311), and incubated for 10 min. The mixture was then added to 2 X 10^5 ESCs seeded in a 6-well plate one day before transfection. Cells were trypsinized and split to three 10-cm plates 16 hours after transfection. 48 post-transfection, medium was added with 2 μg/mL puromycin for selection of clones with gRNA and spCas9 expression. Medium was switched to normal one 96 hours after transfection, and about ten days post-

transfection, colonies were picked by pipette with p200 tips, transferred to 96-well plate, cultured for three days until cells in most of the wells reached full confluence. The 96-well plate was replicated into two new plates, with one plate later used for genomic DNA extraction and PCR amplification to determine the positive clones by PCR primers that target the C-terminus of *Cebpz* and the 6×His-3×Flag-AID sequence (Table 3.1). Once positive clones were screened, corresponding clones in another 96-well plate were transferred to 24-well plates then 6-well plates. Clones were lysed and analyzed by Western blotting using an antibody that recognized CEBPZ to further confirm the addition of tags and similar CEBPZ expression levels as in WT mESCs.

Immunofluorescence staining. Details for the experiment are described in Chapter II. The dilution for primary antibodies were: CTCF (1:500), H3K4me3 (1:250), Fibrillarin (1:250), Flag Tag (1:500). For IAA treatment, mESCs were incubated with 500 μM of IAA for 24 hours and removed before fixation. IF were performed two to three independent times per antibody.

DRIP-qPCR. DNA was extracted from mESCs. Briefly, cells were resuspended in 400 μL ESC lysis buffer (10 mM Tris-Cl pH 7.5, 10 mM EDTA, 10 mM NaCl, 0.5% sarkosyl) with 3.2 μL 20 mg/mL Proteinase K Solution (Bioline, BIO-37084), and incubated at 37 °C overnight. The next day, Phenol/Chloroform extraction was performed, the supernatant was mixed with NaOAc pH 5.2 to 0.3 M, then 2.5 volumes of cold ethanol, kept at -80 °C for 30 min. Spin at maximum speed for 30 min at 4 °C, DNA pellet was washed twice with 70 % ethanol and resuspended in

200 µL mH_2O. DNA concentration was measured by Nanodrop. 150 µg of DNA was mixed with mH_2O to a final volume of 50 µL then with 500 µL MeDIP buffer (10 mM Na_2HPO_4, 140 mM NaCl, 0.05% Triton X-100), sonicated using a Bioruptor (Diagenode) for 15 cycles, 30s on/30s off with the intensity set at medium. 15 µL sonicated samples were as input. 500 µL sonicated samples were mixed with 10 µg mIgG2a antibody or 10 µg S9.6 antibody and rotated 4 °C overnight. The next day, samples were mixed with 50 µL Protein G magnetic beads (NEB, S1430S) pre-washed by MeDIP buffer, rotated at 4 °C for 3 hours. The mixture was then washed with MeDIP buffer 4 X 5 min. To elute, beads were resuspended in 100 µL 2 X STOP buffer (20 mM Tris-Cl pH 8.0, 100 mM NaCl, 20 mM EDTA, 1% SDS), incubated on a Eppendorf ThermoMixer 65 °C 1,000 rpm for 15 min. Supernatant was saved, elution was repeated and combined with the previous one. To input, 2 X STOP buffer was added to 200 µL then both input and elution were added with 200 µL TE buffer (10 mM Tris-Cl pH 8.0, 1 mM EDTA pH 8.0), 4 µL RNase A (Thermo Fisher Scientific, EN0531), 37 °C 1 hour, then 4 µL 20 mg/mL Proteinase K, 55 °C 1 hour. After the treatment, Phenol/Chloroform extraction was performed as described above, DNA was precipitated with ethanol at -20 °C overnight. Next day, DNA was centrifuged at 16,000 g X 45 min at 4 °C and resuspended in 30 µL mH_2O. qPCR was performed using 1 µL of input or elution, three technical replicates each. Primers targeting R-loop-containing regions are listed in Table 3.1.

CUT&RUN. 0.5 million mESCs were resuspended in 1 mL NE buffer (20 mM HEPES-KOH pH 7.9, 10 mM KCl, 0.5 mM Spermidine, 0.1% Triton X-100, 20%

glycerol, 1 × Halt Protease Inhibitor Cocktail (Thermo Fisher Scientific, 78429)), spin down, and resuspended in 600 µL NE buffer. 200 µL BioMag Plus Concanavalin A beads (Polysciences, 86057-10) were washed twice in 1 mL Binding buffer (20 mM HEPES-KOH pH7.9, 10 mM KCl, 1 mM $CaCl_2$, 1 mM $MnCl_2$) and resuspended in 300 µL Binding buffer. Nuclei in NE buffer were gently vortexed and mixed with beads, then rotated 10 min at room temperature. Beads/nuclei mixer were resuspended in 1 mL Blocking buffer (20 mM HEPES pH 7.5, 150 mM NaCl, 0.5 mM Spermidine, 0.1% BSA, 2 mM EDTA, protease inhibitor) and incubated at room temperature for 5 min. Washed once with Wash buffer (same as Blocking buffer without 2 mM EDTA), resuspended in 250 µL Wash buffer, mixed with 250 µL Wash buffer with 5 µL antibody derived from rabbit while vortexing, or 250 µL Wash buffer without antibody as a no antibody negative control. Rotated for 2 hours at 4 °C, then washed twice in 1 mL Wash buffer. Beads/Nuclei mixture were resuspended in 250 µL Wash buffer, while under gentle mixing, 250 µL Wash buffer that contained 2.5 µL protein A-MNase were added. Rotated for 1 hour at 4 °C, then washed twice in 1 mL Wash buffer. Beads/Nuclei mixture were resuspended in 150 µL Wash buffer, left on top of ice water for more than 10 min. To initiate MNase cutting, 3 µL of 100 mM $CaCl_2$ was added to the 150 uL mixture while vortexing, then the mixture was returned to ice water to be incubated for 5 min. The reaction was terminated by addition of 150 µL 2 X STOP buffer (200 mM NaCl, 0.5 mM EDTA, 4 mM EGTA, with 50 µg/mL RNase A, 40 µg/mL glycogen, 10 pg/mL yeast spike-in genomic DNA added freshly) while

vortexing. The reaction was incubated at 37 °C for 20 min, spined 16,000 g for 5 min at 4 °C, supernatant was transferred to a fresh tube. Then 3 µL 10% SDS and 2.5 µL of 20 mg/mL Proteinase K was added and incubated for 10 min at 70 °C. Phenol/Chloroform extraction was performed, supernatant was added with 2 µL of 20 mg/mL glycogen and 750 µL 100% ethanol, mixed, kept at -80 °C for 30 min. Sample was centrifuged at 16,000 g for 45 min at 4 °C. Pellet was washed once in 1 mL 100% ethanol, left dry, and resuspended in 36.5 µL 1 X TE buffer as the fragmented target DNA from the CUT&RUN experiment.

The harvested DNA was then turned into libraries for sequencing. The 36.5 µL DNA was mixed with 5 µL 10 X T4 DNA ligase buffer (NEB, B0202S), 2.5 µL 10 mM dNTPs (NEB), 1.25 mM 10 mM ATP (NEB, P0756S), 3.13 µL 40% PEG 4000, 0.63 µL T4 PNK (NEB, M0201S), 0.5 µL 1:20 diluted T4 DNA polymerase (NEB, M0203S), 0.5 µL Taq DNA polymerase. The mixture was incubated at 12 °C for 15 min, 37 °C for 15 min, and 72 °C for 20 min. TruSeq Universal Adapter was mixed and annealed with each individual TruSeq Indexing Adapter to a final concentration of 1.5 µM. Then 5 µL of each 1.5 µM annealed adapter was mixed with the above 50 µL prepared sample, together with 55 µL Quick Ligation Reaction Buffer (NEB, B2200S) and 5 µL Quick Ligase (NEB, M2200S), incubated at 20 °C for 15 min. The DNA was purified by 38 µL AMpure XP beads (Beckman Coulter, A63881), eluted in 20 µL 10 mM Tris-Cl pH 8.0. The purified DNA was then mixed with 10 µL of 5 X KAPA HiFi GC Buffer (Kapa Biosystems, KK2502), 1.5 µL 10 mM dNTPs, 20 µM TruSeq PE1.0 and PE 2.0 PCR primer, each with a

volume of 5 µL, 7.5 µL mH_2O, 1 µL KAPA HiFi HotStart DNA Polymerase (Kapa Biosystems, KK2502). Amplification was carried out using the following program: 1) 98 °C 45s, 2) 98 °C 15s, 3) 60 °C 10s, 4) Repeat step 2) and 3) for a total of 13-14 cycles, 5) 72 °C 1 min, 6) 4 °C ∞. The amplified library was run in 1.5 % agarose gel, cut at the band size from 150-650 bp, gel purified, and eluted in 30 µL mH_2O. The libraries were mixed at the same amount of DNA, diluted to 1.8 pM, and loaded onto Illumina NextSeq 550 for sequencing.

ChIP-seq. 50 µL Protein A magnetic beads (NEB, S1425S) were washed with PBS with 5 mg/mL BSA three times, then resuspended in 500 µL PBS/BSA. 5 µg primary antibody was added and rotated at 4 °C overnight. The next day, 10 million cells were resuspended in 10 mL PBS and mixed with 1 mL fix solution (11% formaldehyde, 100 mM NaCl, 1 mM EDTA, 50 mM HEPES-KOH pH 7.6), incubated for 10 min at room temperature. 500 µL 2.5M glycine was added and incubated for 5 min at room temperature. Spin 5,000 rpm for 5 min, cell pellet was washed once with PBS + protease inhibitor, then resuspended in 640 µL SDS lysis buffer (1% SDS, 10 mM EDTA, 50 mM Tris-Cl pH 8.0, protease inhibitor) and incubated on ice for 10 min. Cell pellet was sonicated using a Bioruptor for 30 cycles, 30s on/30s off with the intensity set at high, spin down 16,000g for 10 min at 4 °C. 30 µL sample was saved as input, the rest was added with 2 mL IP buffer (0.01% SDS, 1.1% Triton X-100, 1.2 mM EDTA, 16.7 mM Tris-Cl pH 8.0, 167 mM NaCl, protease inhibitor). 1.25 mL sample was mixed with antibody/beads and rotated 4 °C overnight. The next day, washed twice in IP buffer, and five times in

MVL buffer (50 mM Tris-Cl pH7.4, 250 mM NaCl, 1 mM EDTA, 0.1% Triton X-100, protease inhibitor), 5 min each at 4 °C. To elute, beads were resuspended in 100 µL 2 X STOP buffer (20 mM Tris-Cl pH 8.0, 100 mM NaCl, 20 mM EDTA, 1% SDS), incubated on a Eppendorf ThermoMixer 65 °C 1,000 rpm for 15 min. Supernatant was saved, elution was repeated and combined with the previous one. To input, 170 µL 2 X STOP buffer was added, then elution and input were incubated at 65 °C overnight. The next day, RNase A treatment, protease K treatment, DNA extraction was carried out as described in the DRIP-qPCR procedure, DNA pellet was resuspended in 40 µL mH2O. Library preparation and sequencing steps were similar to what was described in the CUT&RUN procedure.

ATAC-seq. 50,000 mESCs were washed with PBS and resuspended in 50 µL Lysis buffer (10 mM Tris-Cl pH 7.5, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% NP-40, 0.1% Tween-20, 0.01% Digitonin) and incubated on ice for 15 min. The cells were resuspended in wash buffer (10 mM Tris-Cl pH 7.5, 10 mM NaCl, 3 mM $MgCl_2$-,0.1% Tween-20), spin down 500 g for 10 min at 4 °C. The nuclei pellet was mixed with 25 µL Tagment DNA buffer (Illumina, 15027866), 16.5 µL PBS, 0.5 µL 10% Tween-20, 0.5 µL 1% Digitonin, 2.5 µL Tagment DNA Enzyme 1 (Illumina, 1027865), 5 µL mH_2O, incubated at 37 °C for 30 min with vortexing at 1,000 rpm. DNA was isolated using MinElute Reaction Cleanup Kit (Qiagen, 28204), eluted in 10 µL EB. The purified transposed DNA was mixed with 10 µL mH_2O, 2.5 µL 25 µM Nextera S50* barcoded primer, 2.5 µL 25 µM Nextera N70* barcoded primer, 25 µL NEBNext High-Fidelity 2X PCR Master Mix (NEB, M0541S). PCR

amplification was carried out using: 1) 72 °C 5 min, 2) 98 °C 30 s, 3) 98 °C 10 s, 4) 63 °C 30 s, 5) 72 °C 1 min, 6) Repeated step 3) 4) and 5) for a total of 10 cycles, 7) 72 °C 3 min, 8) 4 °C ∞. Sample was run in 1.5% agarose gel with the band at the size of 150 – 500 bp cut and gel purified. DNA was eluted in 20 µL mH_2O. Libraries were mixed at an equal amount and loaded for sequencing.

Western blotting. Dilution for antibodies was: CTCF (1:4,000), CEBPZ (1:2,000), β-actin (1:2,000).

RT-qPCR. RNA was extracted using an RNA Clean & Concentrator-25 Kit (Zymo Research, 11-353B) with on-column DNase I digestion for 1 hour. cDNA was synthesized using purified MMTV reverse transcriptase. Using primers that target cDNA, quantification was performed with *Gapdh* used as a loading control (Table 3.1).

Data availability. CUT&RUN, ChIP-seq, ATAC-seq data are deposited in Gene Expression Omnibus (GSE185181).

Table 3.1. CRISPR gRNAs that target *Cebpz*. Primers for performing DRIP-qPCR, and RT-qPCR

Primer sets	Sequence
Cebpztag_gRNA_F	CACCGATGTCACTTCCTCTGCCGTT
Cebpztag_gRNA_R	AAACAACGGCAGAGGAAGTGACATC
Cebpztag_inner_F	TGACTGGCTGCACAACAGAGA
Cebpztag_inner_R	AGTCTGGAATGTATGGAGTTAGAGAGA
Cebpztag_outer_F	ATAGCCTTGCCCACCCTTCC
Cebpztag_outer_R	CAGATTTCAGACAAACAATGAGCAAGA
Wipf2_R-loop_F	GATCCATTTCCGGGTTGGTAA
Wipf2_R-loop_R	TTTAGTCCTGCTCGTTCGCC
Sp2_R-loop_F	GATCGCTGTGAGTGTGAGGCTAA
Sp2_R-loop_R	TCTTCCTCTCTTTGTTGTTGTTGACT
Jarid2_R-loop_F	GGCGTGACTCTAACTAAGGAGGTG
Jarid2_R-loop_R	GCTGCGGGATGAACCGAACG
Gphn_R-loop_F	CAGTACGAATACAGACCGTGAAAGC
Gphn_R-loop_R	CACAAGCCAGTTATCCCTGTGGTA
5ETS_F	CTTGCGTGTGCTTGCTGT
5ETS_R	GAAATCGGGAAAAACGTCTG
18S_F	CTATCAACTTTCGATGGTAGTCGCC
18S_R	CTTGGATGTGGTAGCCGTTTCTC
28S_F	CTAGCAGCCGACTTAGAACTGGT
28S_R	CAGAAATCACATCGCGTCAACACC
Gapdh_RTqPCR_F	TTGATGGCAACAATCTCCAC
Gapdh_RTqPCR_R	CGTCCCGTAGACAAAATGGT

Table 3.2. gBlock dsDNA that targets *Cebpz* for adding AID tag

Name	Sequence
6×His-3×Flag-AID	gtaagtaagctggtggtgtttgtttgtttgtttgtttgtgttgtttttacagcagggtcgtagcccaggctg accttgggctctaaatcctcctcctcccaagcagtaggattacaagtgcactaccaggtcacctgt gtcatttaaatggtacttccctgaagatgtccttagtggcaggatgtttgtcacatagcaaggcattt aaaccacacacacaaactgcagacccacacacgttatttcttcaagagtctgtctgttttatttaga

	cttagctaaaggtgtactctcacatgcttgaagactttgtaattattcctgatgttataatcactagtttc tattctttcataaaatagaaaagaaaaatctttccagtataaatcctatattttcacttgtaaattcctc aactatccagtagaggagagcagcaaacagtgggtgggaagccgcaggcctgcagttacgg tcagggagcgtctgtggagtcccgggtacagactcatctgtcgtctcccctcaaagGTTTCA AACAGCTCAAGTGGGAAGCTGAGCGAGATGACTGGCTGCACAAC AGAGATGTGAAAAGTATCATCAAGAAAAAGAAAAATTTCAGGAAG AAGATGAAAGCCCCTCAGAAACCgAAACGGCAGAGGAAGtccggaa gaggatcgCATCACCACCATCATCACGCTGGCGCAGACTACAAAGA CCATGACGGTGATTATAAAGATCATGACATCGACTACAAGGATGA CGATGACAAGggaggctctgggagcggaCCAAAGGACCCAGCTAAGCCT CCCGCTAAAGCCCAGGTGGTCGGATGGCCACCTGTCCGATCCTA TCGGAAGAATGTGATGGTGTCTTGCCAGAAGAGCTCTGGCGGAC CAGAGGCTTGACATCTTCAGGGTGACAATAAATTAAGATTATACT TACCCTTAGTTTTTGCTAGTCAACCATTTTTCCTCATCTTTCTCTC TAACTCCATACATTCCAGACTGTTCAATGGATTTGTAATAAACTGT GGAAACAAAGTACTGTCATTTATAAAATTACACAAAATTTAAACTT ATAGTGAAAGAATTCTTCCATATTGCTACATAGCAAACGAAAGGA CAGCGGGGCTCCATGTCCACACGGGAGCTGTAACCACTCTCCA GACCTCTGTGATGGGCTGCTTGTCCGGGACCAGGTACCATGCTT AGGCTAAGTGCATGATGTCACTGTGGTCAACTCTAACTATCAGG GTCAGCGTTGGCTCTCCCCAAGAGAAAGCAGTATCTCTGAGAAC ATAGTACACCGGAGGAAAACAGTCTTTACAGCAGTAACTATGGTT TGTGCCCGTCCCAACCCTTAGGTCTCGCTCTGGAATCTATAATCT TAACCACTGAACGAATGAAAACTCCAGCTGGCACGTGGGGTATA TAACTCAGTGACAGCACTTTTGTCCATTGTTTATAAAGCCTGCAG GTTGATCAGTAAACAGGGGAATAAGGATTAAAACAA

CHAPTER IV: Discussion and future directions

Summary

RNA can associate with dsDNA in a sequence-dependent manner to form RNA-DNA hybrids, with ssDNA corresponding to the non-complementary DNA strand looped out. This three-stranded nucleic acid structure is known as an R-loop. R-loops are usually formed during transcription when the RNA transcribed by RNAP threads back and anneal with the ssDNA it transcribed from. The primary interest of researchers focused on R-loops has been R-loops associated with pre-mRNAs at coding genes. However, R-loops have also been reported at tRNA genes transcribed by RNAP III (El Hage *et al.*, 2014; Liu and Sun, 2021) and rRNA loci transcribed by RNAP I (El Hage *et al.*, 2010; Velichko *et al.*, 2019). Even though these classes of R-loops are less well characterized, rRNA-associated R-loops, in particular, appear to be a large source of R-loops found in cells, due to the high level of rRNA transcription. Therefore, it is clear that greater attention should be focused on R-loops at the rDNA loci in studies of R-loop functions.

R-loops have been shown to regulate transcription initiation and termination, class-switch recombination, DNA repair, and other processes. Moreover, the R-loops that accumulate to high levels in the genome have long been known as a source of genomic instability, indicating that R-loops need to be tightly regulated. Indeed, since the discovery of R-loops, various factors, including ribonucleases, RNA helicases, splicing factors, and topoisomerases, have been

shown to associate with and regulate R-loop abundance. The aim of this thesis has been to identify R-loop-interacting proteins in a systematic way. In the first part of the thesis, I described a method using an antibody that recognizes R-loops to capture R-loops in conjunction with their interacting proteins. I found that a large fraction of the R-loop interactome consists of proteins either partially or completely localized to the nucleolus. I examined several identified DEAD-box helicases and uncovered important regulatory roles in rRNA processing. In the second part of the thesis, I characterized a transcription factor named CEBPZ that is enriched on or near R-loops. I found that CEBPZ regulates mRNA- and rRNA-associated R-loops, and associates with the insulator-binding protein CTCF. In this chapter, I will highlight key features of my findings described in Chapters II and III, delve into how these findings contribute to our understanding of R-loops, and propose future directions that may further advance this field.

Interplay between rRNA synthesis and R-loops

Using high salt washes to remove nuclear proteins that non-specifically associate with R-loop-proximal chromatin, followed by chromatin fragmentation and R-loop capture by S9.6 antibody, I identified a stringent set of R-loop-associated proteins. By introducing 4SU-dependent crosslinking before the IP steps, proteins that directly interact with the RNA part of R-loop could be captured more efficiently, enabling identification of additional R-loop-binding proteins.

Several proteins known to regulate R-loops, including RNA helicases DHX9 and DDX21, along with numerous additional RNA-binding proteins, were identified, validating these approaches. It was noticeable that proteins with no known functions in regulation of R-loops but known functions in pre-rRNA processing and ribosome biogenesis, such as SPB1 and NOG1, were highly enriched by S9.6. Functional analysis of the R-loop-associated proteins indicated rRNA processing factors were highly overrepresented. By overlapping the R-loop interactome with the nucleolar proteome of mouse fibroblasts, I observed numerous proteins shared between the two data sets, further confirming that many of the most highly enriched R-loop-associated proteins localize mainly to the nucleolus. Known as co-transcriptionally formed structures, R-loops at coding genes have been studied most intensively, with rRNA- and tRNA-dependent R-loops only occasionally mentioned in much of the previous literature (El Hage *et al.*, 2010, 2014). However, my findings suggest that R-loops formed during rRNA transcription and processing should not be ignored. In addition to my findings that R-loop binding proteins are largely nucleolar, IF using the S9.6 RNA/DNA hybrid antibody shows strong staining in nucleoli (García-Rubio *et al.*, 2015; Shen *et al.*, 2017). Moreover, inhibition of rRNA transcription abolished nucleolar R-loops detected by IF, causing loss of nucleolar localization of a significant portion of RNase H1 and Top1 (Shen *et al.*, 2017). Together, my studies and previous studies indicate that rRNA-dependent R-loops contribute to a high percent of R-loops in cells and lead to R-loop-associated factors being identified within nucleoli.

RNA helicases belonging to the DEAD-box family were highly enriched in R-loop immunoprecipitates, with DDX5 and DDX21 already identified as R-loop regulators (Song *et al.*, 2017; Mersaoui *et al.*, 2018). DEAD-box helicases are known to play roles in mRNA transcription, splicing, rRNA processing, and ribosome biogenesis (Rocak and Linder, 2004; Martin *et al.*, 2013). IF of DDX18, DDX24, and DDX27 indicated their strong nucleolar localization and weaker staining in nuclear regions outside of nucleoli. Northern blotting using probes targeting 18S and pre-rRNAs that contain the 18S sequence indicated that DDX10 and DDX24 function at different rRNA processing steps. This agrees with the previous research that DDX10 and DDX24 affect pre-rRNAs that lead to mature 18S (Zagulski *et al.*, 2003; Turner *et al.*, 2009; Yamauchi *et al.*, 2014), and adds an additional layer of mechanistic insight. Similarly, by performing Northern blotting using a probe that targets 28S and pre-rRNA that contain the 28S region, I found that DDX27 and DDX54 contribute to 28S processing.

CEBPZ, which was identified by the uncrosslinked and crosslinked S9.6 co-IP, is a transcription factor previously shown to drive the expression of the *hsp* promoter (Lum *et al.*, 1990; Imbriano *et al.*, 2001). However, IF using antibodies that target endogenously epitope-tagged CEBPZ revealed strong nucleolar staining. Rapid depletion of CEBPZ caused downregulation of 18S-containing and 28S-containing rRNAs but not 47S pre-rRNA, the latter of which corresponds to unprocessed nascent pre-rRNA, indicating that CEBPZ regulates some steps of rRNA processing but not rRNA transcription. In agreement with this observation,

depletion of CEBPZ induced downregulation of R-loops associated with 18S and 28S regions, possibly due to mis-processing of the underlying pre-rRNAs. This observation supports the argument that by affecting the abundance and dynamics of rRNA processing in nucleoli, the rRNA regulators and nucleolar proteins characterized in this thesis affect the abundance of R-loops in cells.

It is worth noting that IF of S9.6 was recently shown to exhibit a bias toward certain RNA species, potentially due to low-affinity dsRNA binding (Phillips *et al.*, 2013; Hartono *et al.*, 2018). Indeed, Smolka *et al.* showed that S9.6 has a strong affinity for rRNAs by EMSA, contributing to the strong nucleolar and cytoplasmic staining by IF (Smolka *et al.*, 2021). By performing RNase A and DNase I treatment, I was able to show that the ability of S9.6 to co-IP proteins was dependent on both RNA and DNA. Although this experiment rules out the possibility that S9.6 binds to free rRNAs to capture a bunch of R-loop-independent rRNA-interacting proteins, the R-loops interactomes should still be interpreted carefully. A mixture of RNase III and RNase T1 treatment, which digests dsRNA and ssRNA under defined conditions, could be carried out prior to the S9.6 IP to eliminate the interference of dsRNAs and rRNAs. Once R-loop-interacting proteins are identified, DRIP-qPCR or DRIP-seq, which are more robust than S9.6 IF (Smolka *et al.*, 2021), should be carried as proof that the identified R-loop-interacting proteins actually regulate R-loop abundance.

Functions of nucleolar proteins outside of the nucleolus

IF of several DEAD-box helicases showed strong nucleolar localization but also some degree of nuclear staining in regions other than the nucleolus, raising the possibility these proteins function in regulation of mRNA expression. mRNA-seq upon *Ddx10*, *Ddx24*, *Ddx27*, *Ddx54* KD shows that a shared set of genes were misregulated upon the four KDs, including genes associated with development and differentiation. The fact that largely the same genes were affected upon KD of each of the four *Ddx* genes could reflect a shared set of direct mRNA targets. Alternatively, these findings could indicate that the genes misregulated in each KD are more sensitive to alterations in rRNA levels.

It has been observed for some DEAD-box proteins with strong nucleolar localization, that they can function outside of nucleoli. DDX21 and DDX5 were shown by IF and nuclear fractionation to be nucleolar proteins, with known functions in regulating rRNA transcription and processing (Saporita *et al.*, 2011; Song *et al.*, 2017). They were also shown to regulate transcription of mRNAs: DDX21 was shown to resolve R-loops formed due to RNAP II pausing and affect expression of the associated genes (Song *et al.*, 2017), while DDX5 has been shown to interact with the transcription factor Fra-1 and help regulate the expression of Fra-1 target genes (He *et al.*, 2019). To investigate the possibility that DDX10, DDX24, DDX27, and DDX54 regulate a set of genes through a shared pathway, genome localization of the DEAD-box proteins could be studied by ChIP-seq or CUT&RUN to identify shared target genes. In addition, proteomics

approaches could be performed to reveal the transcription factors or other proteins that bind to the four DEAD-box proteins. Furthermore, mRNA-seq could be carried out upon treatment with drugs that disrupt rRNA processing or inhibit rRNA transcription to study if the same genes misregulated upon KD of the four DEAD-box helicases are also misregulated upon chemical inhibition of rRNA expression.

CEBPZ, a transcription factor that drives the expression of *hsp* genes (Lum *et al.*, 1990; Imbriano *et al.*, 2001), is also known to regulate the recruitment of METTL3 (Barbieri *et al.*, 2017). METTL3 catalyzes deposition of m6A on RNAs, which in turn favors formation of R-loops (Barbieri *et al.*, 2017). CEBPZ CUT&RUN showed binding to known targets as well as new target genes. I found that depletion of CEBPZ caused a reduction in R-loops that localize close to CEBPZ binding sites. However, unexpected nucleolar localization was observed in CEBPZ IF studies. Depletion of CEBPZ affected processing but not transcription of rRNA. Moreover, reduction of R-loops at rDNA regions was observed upon CEBPZ depletion. In total, these results indicate that CEBPZ is a protein with largely nucleolar distribution and functions, which also has roles in regulating mRNA expression and its associated R-loops.

Comparison of R-loop-interactomes from different studies

At this time, several groups have published different methods to identify R-loop-interacting proteins. Among them, the Gromak group used the S9.6 antibody

to capture R-loops and binding proteins from HeLa cells (Cristini *et al.*, 2018). A similar technique was used by another group to study the R-loop-interactome in mESCs, with 229 proteins identified (Li *et al.*, 2020). In contrast to my approach, neither of these studies used a stringent wash step to remove proteins loosely bound to R-loops or R-loop proximal chromatin. In addition, no previous approach used selective RNA-protein crosslinking to stabilize weakly binding proteins that associate with the RNA component of R-loops. Instead of using S9.6 antibody to capture existing R-loops in cells, Wang *et al.* used two different *in vitro* generated RDHs (produced by annealing RNA and DNA oligonucleotides), whose sequences were derived from the genomic regions known to form R-loops in cells. Using these RDHs to pull down proteins from human B-cell extracts, they were able to identify 803 proteins shared between the two (Wang *et al.*, 2018). By comparing the 364 proteins identified by our uncrosslinked and crosslinked S9.6 co-IP study and the 229 proteins from Li *et al.*, both of which were performed in mESCs, 123 proteins were identified that were shared by the two studies (Observed/expected = 29.5; p-value = 6.366×10^{-158}). The high number of shared proteins observed between the two studies suggests that many R-loop binding proteins are insensitive to moderate differences in the methods. The proteins that are specific to one study or the other may result from differences in the methods, including the methods of extracting nuclei, the presence or absence of crosslinking, different components of wash and IP buffers, and other differences. By comparing the 364 proteins identified in my studies with the 469 proteins from Cristini *et al.*, in which mouse to

human homologs could be identified, 172 proteins were found to be shared (O/E = 20.5, p-value = 4.346×10^{-191}). Similarly, 156 overlapping proteins were found in between our study and the 803 proteins from Wang *et al.* (O/E = 10.7, p-value = 1.006×10^{-121}). The lower but still high observed/expected ratio between our study and Wang *et al.* could be explained by their use of annealed RDHs as probes rather than use of S9.6 to capture R-loops. Moreover, by overlapping the lists of proteins identified from the four studies, a shared of 36 proteins can still be identified, including known R-loop regulators such as DHX9, DDX21, and NONO.

R-loops and other non-canonical nucleic acid structures, such as G4s, have previously been shown to colocalize in the genome (De Magis *et al.*, 2019). By comparing 77 known G4-binding proteins from *Homo sapiens* (Brázda *et al.*, 2018) with the 364 R-loop-interacting proteins from this thesis, 21 of the 77 G4-binding proteins overlap with the 364 mESC R-loop-binding proteins (O/E = 15.0; p-value = 2.572×10^{-19}). One overlapping protein is the DDX21 helicase, consistent with studies showing that DDX21 resolves R-loops and G4s (McRae *et al.*, 2017; Song *et al.*, 2017). m6A modification has been shown to accumulate on the RNA components of R-loops and favor R-loop formation (Abakir *et al.*, 2019; Yang *et al.*, 2019). Comparison of the R-loop interactome identified in this thesis with the 915 m6A-interacting proteins previously identified in mESCs (Edupuganti *et al.*, 2017) revealed 229 shared proteins (O/E = 13.8, p-value = 1.062×10^{-219}). Among the overlapping proteins were CEBPZ and HNRNPA2B1, the latter of which is

known as an m6A reader and has been shown to induce R-loop-dependent DNA damage (Alarcón *et al.*, 2015; Abakir *et al.*, 2019).

Significant effects of CEBPZ on CUT&RUN performance

To further study the roles of CEBPZ in regulation of R-loops, I performed CUT&RUN to map its genomic distribution. Depletion of CEBPZ followed by DRIP-qPCR showed that CEBPZ regulates the abundance of nearby R-loops but not R-loops far from CEBPZ binding sites. When performing motif analysis for CEBPZ CUT&RUN, the CTCF binding motif was highly ranked, consistent with my finding that CTCF was enriched by S9.6 and with previous research that CTCF co-localizes with R-loops. Motif analysis of CTCF CUT&RUN also revealed the CEBPZ binding motif, and co-IP experiment confirmed their physical interaction in the nucleus.

Though both were identified as R-loop-binding proteins, CEBPZ and CTCF were not shown to bind at shared genomic regions and were not previously observed to interact in proteomics studies (Marino *et al.*, 2019; Lehman *et al.*, 2021). To investigate if CEBPZ could function to recruit CTCF and/or regulate CTCF binding at shared CEBPZ/CTCF binding sites, I performed CTCF CUT&RUN upon depletion of CEBPZ, which showed a significant decrease of CTCF binding over its binding sites upon CEBPZ depletion. However, the reduction of CTCF binding occurred not only at the shared binding sites with CEBPZ but also at CTCF-only sites, suggesting the possibility that CEBPZ

depletion may non-specifically affect the performance of CUT&RUN. To further examine this possibility, I performed CTCF ChIP-seq upon CEBPZ depletion, which showed no reduction of CTCF binding at its binding sites. Moreover, by performing CUT&RUN of H3K4me3, which has not been reported to be affected by CEBPZ, H3K4me3 distribution around TSSs was also dramatically decreased upon CEBPZ depletion, confirming that CEBPZ depletion affects CUT&RUN performance. CEBPZ depletion resulted in higher levels of DNA released by CUT&RUN, independent of whether primary antibodies were included to recruit protein A-MNase to specific loci. To figure out the mechanisms by which CEBPZ affects CUT&RUN performance, I performed ATAC-seq to look for chromatin accessibility change upon CEBPZ depletion, which showed little change. Similarly, IF of CTCF, H3K4me3, and DAPI staining revealed no change in CTCF and H3K4me3 nuclear distribution, and little change of nuclear morphology upon CEBPZ depletion.

Further studies need to be carried out to understand the mechanism by which CEBPZ depletion affects CUT&RUN performance. One possibility is that DNA damage is increased upon CEBPZ loss, causing additional DNA to be liberated from bulk chromatin and included in CUT&RUN libraries. This possibility could be tested by looking for DNA lesions upon CEBPZ depletion, as well as performing a “mock” CUT&RUN experiment in which protein A-MNase was left out in cells with or without CEBPZ depletion. Another direction would be to look for changes in nucleosome positioning or density, which may affect the digestion

kinetics of protein A-MNase in background regions of the genome. Investigation of how CUT&RUN is impacted by CEBPZ loss may help us better understand the roles of CEBPZ in regulation of global chromatin structure or genome integrity, adding to its known roles in regulation of rRNA processing, R-loops, m6A deposition, and transcription.

BIBLIOGRAPHY

Abakir, A. *et al.* (2019) 'N 6-methyladenosine regulates the stability of RNA:DNA hybrids in human cells', *Nature Genetics*. Nature Research. doi: 10.1038/s41588-019-0549-x.

Agoff, S. N. *et al.* (1993) 'Regulation of the human hsp70 promoter by p53', *Science*. doi: 10.1126/science.8418500.

Agoff, S. N. and Wu, B. (1994) 'CBF mediates adenovirus Ela trans-activation by interaction at the C-terminal promoter targeting domain of conserved region 3', *Oncogene*.

Aguilera, A. and García-Muse, T. (2012) 'R Loops: From Transcription Byproducts to Threats to Genome Stability', *Molecular Cell*. doi: 10.1016/j.molcel.2012.04.009.

Alarcón, C. R. *et al.* (2015) 'HNRNPA2B1 Is a Mediator of m6A-Dependent Nuclear RNA Processing Events', *Cell*. doi: 10.1016/j.cell.2015.08.011.

Allison, D. F. and Wang, G. G. (2019) 'R-loops: Formation, function, and relevance to cell stress', *Cell Stress*. doi: 10.15698/cst2019.02.175.

Amano, T. *et al.* (2009) 'Chromosomal Dynamics at the Shh Locus: Limb Bud-Specific Differential Regulation of Competence and Active Transcription', *Developmental Cell*. doi: 10.1016/j.devcel.2008.11.011.

Arab, K. *et al.* (2019) 'GADD45A binds R-loops and recruits TET1 to CpG island

promoters', *Nature Genetics*. doi: 10.1038/s41588-018-0306-6.

Bacolla, A., Wang, G. and Vasquez, K. M. (2015) 'New Perspectives on DNA and RNA Triplexes As Effectors of Biological Activity', *PLoS Genetics*. doi: 10.1371/journal.pgen.1005696.

Baranello, L. *et al.* (2009) 'DNA topoisomerase I inhibition by camptothecin induces escape of RNA polymerase II from promoter-proximal pause site, antisense transcription and histone acetylation at the human HIF-1α gene locus', *Nucleic Acids Research*. doi: 10.1093/nar/gkp817.

Baranello, L. *et al.* (2016) 'ChIP bias as a function of cross-linking time', *Chromosome Research*. doi: 10.1007/s10577-015-9509-1.

Barbieri, I. *et al.* (2017) 'Promoter-bound METTL3 maintains myeloid leukaemia by m6A-dependent translation control', *Nature*. Nature Publishing Group, 552(7683), pp. 126–131. doi: 10.1038/nature24678.

Beckedorff, F. C. *et al.* (2013) 'The Intronic Long Noncoding RNA ANRASSF1 Recruits PRC2 to the RASSF1A Promoter, Reducing the Expression of RASSF1A and Increasing Cell Proliferation', *PLoS Genetics*. doi: 10.1371/journal.pgen.1003705.

Beddington, R. S. P. and Robertson, E. J. (1989) 'An assessment of the developmental potential of embryonic stem cells in the midgestation mouse embryo', *Development*. doi: 10.1242/dev.105.4.733.

Bell, A. C. and Felsenfeld, G. (2000) ‘Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene’, *Nature*. doi: 10.1038/35013100.

Bell, A. C., West, A. G. and Felsenfeld, G. (1999) ‘The protein CTCF is required for the enhancer blocking activity of vertebrate insulators’, *Cell*. doi: 10.1016/S0092-8674(00)81967-4.

Belotserkovskii, B. P. *et al.* (2018) ‘R-loop generation during transcription: Formation, processing and cellular outcomes’, *DNA Repair*. doi: 10.1016/j.dnarep.2018.08.009.

Bentley, D. L. (2014) ‘Coupling mRNA processing with transcription in time and space’, *Nature Reviews Genetics*. doi: 10.1038/nrg3662.

Boguslawski, S. J. *et al.* (1986) ‘Characterization of monoclonal antibody to DNA.RNA and its application to immunodetection of hybrids.’, *Journal of immunological methods*, 89(1), pp. 123–130. doi: 10.1016/0022-1759(86)90040-2.

Bonner, J. *et al.* (1968) ‘The biology of isolated chromatin’, *Science*. doi: 10.1126/science.159.3810.47.

Brázda, V. *et al.* (2018) ‘The amino acid composition of quadruplex binding proteins reveals a shared motif and predicts new potential quadruplex interactors’, *Molecules*. doi: 10.3390/molecules23092341.

Buenrostro, J. D. *et al.* (2015) ‘ATAC-seq: A method for assaying chromatin accessibility genome-wide’, *Current Protocols in Molecular Biology*. doi: 10.1002/0471142727.mb2129s109.

Cartwright, P. *et al.* (2005) ‘LIF/STAT3 controls ES cell self-renewal and pluripotency by a Myc-dependent mechanism’, *Development*. doi: 10.1242/dev.01670.

Cerritelli, S. M. and Crouch, R. J. (2009) ‘Ribonuclease H: the enzymes in eukaryotes.’, *The FEBS journal*. NIH Public Access, 276(6), pp. 1494–505. doi: 10.1111/j.1742-4658.2009.06908.x.

Chae, H. D., Yun, J. and Shin, D. Y. (2005) ‘Transcription repression of a CCAAT-binding transcription factor CBF/HSP70 by p53’, *Experimental and Molecular Medicine*. doi: 10.1038/emm.2005.60.

Chakraborty, P. and Grosse, F. (2011) ‘Human DHX9 helicase preferentially unwinds RNA-containing displacement loops (R-loops) and G-quadruplexes’, *DNA Repair*, 10(6), pp. 654–665. doi: 10.1016/j.dnarep.2011.04.013.

Chakraborty, P., Huang, J. T. J. and Hiom, K. (2018) ‘DHX9 helicase promotes R-loop formation in cells with impaired RNA splicing’, *Nature Communications*. doi: 10.1038/s41467-018-06677-1.

Chalkley, R. J. *et al.* (2005) ‘Comprehensive analysis of a multidimensional liquid chromatography mass spectrometry dataset acquired on a quadrupole selecting,

quadrupole collision cells, time-of-flight mass spectrometer: II New developments in protein prospector allow for reliable and', *Molecular and Cellular Proteomics*. doi: 10.1074/mcp.D500002-MCP200.

Chédin, F. *et al.* (2021) 'Best practices for the visualization, mapping, and manipulation of R-loops', *The EMBO Journal*. doi: 10.15252/embj.2020106394.

Chen, G. (2013) *Characterization of protein therapeutics using mass spectrometry*, *Characterization of Protein Therapeutics using Mass Spectrometry*. doi: 10.1007/978-1-4419-7862-2.

Chen, J. Y. *et al.* (2019) 'R-ChIP for genome-wide mapping of R-loops by using catalytically inactive RNASEH1', *Nature Protocols*. doi: 10.1038/s41596-019-0154-6.

Chen, L. *et al.* (2017) 'R-ChIP Using Inactive RNase H Reveals Dynamic Coupling of R-loops with Transcriptional Pausing at Gene Promoters.', *Molecular cell*. Elsevier, 68(4), pp. 745-757.e5. doi: 10.1016/j.molcel.2017.10.008.

Chen, P. B. *et al.* (2013) 'Hdac6 regulates Tip60-p400 function in stem cells', *eLife*, 2013(2). doi: 10.7554/eLife.01557.

Chen, P. B. *et al.* (2015) 'R loops regulate promoter-proximal chromatin architecture and cellular differentiation.', *Nature structural & molecular biology*, 22(12), pp. 999–1007. doi: 10.1038/nsmb.3122.

Chu, F. *et al.* (2019) 'Allosteric Regulation of Rod Photoreceptor

Phosphodiesterase 6 (PDE6) Elucidated by Chemical Cross-Linking and Quantitative Mass Spectrometry', *Journal of Molecular Biology*. doi: 10.1016/j.jmb.2019.07.035.

Condic, M. L. (2014) 'Totipotency: What it is and what it is not', *Stem Cells and Development*. doi: 10.1089/scd.2013.0364.

Cordin, O. *et al.* (2006) 'The DEAD-box protein family of RNA helicases', *Gene*. doi: 10.1016/j.gene.2005.10.019.

Cornelio, D. A. *et al.* (2017) 'Both R-loop removal and ribonucleotide excision repair activities of RNase H2 contribute substantially to chromosome stability', *DNA Repair*, 52, pp. 110–114. doi: 10.1016/j.dnarep.2017.02.012.

Cristini, A. *et al.* (2018) 'RNA/DNA Hybrid Interactome Identifies DXH9 as a Molecular Player in Transcriptional Termination and R-Loop-Associated DNA Damage.', *Cell reports*. Elsevier, 23(6), pp. 1891–1905. doi: 10.1016/j.celrep.2018.04.025.

Deutsch, E. W. *et al.* (2020) 'The ProteomeXchange consortium in 2020: Enabling "big data" approaches in proteomics', *Nucleic Acids Research*. doi: 10.1093/nar/gkz984.

Dixon, J. R. *et al.* (2012) 'Topological domains in mammalian genomes identified by analysis of chromatin interactions', *Nature*. doi: 10.1038/nature11082.

Domínguez-Sánchez, M. S. *et al.* (2011) 'Genome instability and transcription

elongation impairment in human cells depleted of THO/TREX', *PLoS Genetics*. doi: 10.1371/journal.pgen.1002386.

Duquette, M. L. *et al.* (2004) 'Intracellular transcription of G-rich DNAs induces formation of G-loops, novel structures containing G4 DNA.', *Genes & development*, 18(13), pp. 1618–29. doi: 10.1101/gad.1200804.

Edskes, H. K., Ohtake, Y. and Wickner, R. B. (1998) 'Mak21p of Saccharomyces cerevisiae, a homolog of human CAATT-binding protein, is essential for 60 S ribosomal subunit biogenesis', *Journal of Biological Chemistry*. doi: 10.1074/jbc.273.44.28912.

Edupuganti, R. R. *et al.* (2017) 'N6-methyladenosine (m6A) recruits and repels proteins to regulate mRNA homeostasis', *Nature Structural and Molecular Biology*. doi: 10.1038/nsmb.3462.

Evans, M. J. and Kaufman, M. H. (1981) 'Establishment in culture of pluripotential cells from mouse embryos', *Nature*. doi: 10.1038/292154a0.

Fazzio, T. G., Huff, J. T. and Panning, B. (2008) 'An RNAi Screen of Chromatin Proteins Identifies Tip60-p400 as a Regulator of Embryonic Stem Cell Identity', *Cell*, 134(1), pp. 162–174. doi: 10.1016/j.cell.2008.05.031.

Fong, Y. W. and Zhou, Q. (2001) 'Stimulatory effect of splicing factors on transcriptional elongation', *Nature*. doi: 10.1038/414929a.

Fudenberg, G. *et al.* (2016) 'Formation of Chromosomal Domains by Loop

Extrusion', *Cell Reports*. doi: 10.1016/j.celrep.2016.04.085.

Fütterer, A. *et al.* (2021) 'Impaired stem cell differentiation and somatic cell reprogramming in DIDO3 mutants with altered RNA processing and increased R-loop levels', *Cell Death & Disease*. doi: 10.1038/s41419-021-03906-2.

Gan, W. *et al.* (2011) 'R-loop-mediated genomic instability is caused by impairment of replication fork progression.', *Genes & development*. Cold Spring Harbor Laboratory Press, 25(19), pp. 2041–56. doi: 10.1101/gad.17010011.

García-Rubio, M. L. *et al.* (2015) 'The Fanconi Anemia Pathway Protects Genome Integrity from R-loops', *PLoS Genetics*. doi: 10.1371/journal.pgen.1005674.

Ginno, P. A. *et al.* (2012) 'R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters.', *Molecular cell*. Elsevier, 45(6), pp. 814–25. doi: 10.1016/j.molcel.2012.01.017.

Ginno, P. A. *et al.* (2013) 'GC skew at the 5' and 3' ends of human genes links R-loop formation to epigenetic regulation and transcription termination', *Genome Research*, 23(10), pp. 1590–1600. doi: 10.1101/gr.158436.113.

Gómez-González, B. *et al.* (2011) 'Genome-wide function of THO/TREX in active genes prevents R-loop-dependent replication obstacles', *EMBO Journal*. doi: 10.1038/emboj.2011.206.

Gómez-González, B. and Aguilera, A. (2007) 'Activation-induced cytidine

deaminase action is strongly stimulated by mutations of the THO complex', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.0702836104.

Goñi, J. R., de la Cruz, X. and Orozco, M. (2004) 'Triplex-forming oligonucleotide target sequences in the human genome', *Nucleic Acids Research*. doi: 10.1093/nar/gkh188.

Grote, P. and Herrmann, B. G. (2013) 'The long non-coding RNA Fendrr links epigenetic control mechanisms to gene regulatory networks in mammalian embryogenesis', *RNA Biology*. doi: 10.4161/rna.26165.

Gyi, J. I. *et al.* (1996) 'Comparison of the thermodynamic stabilities and solution conformations of DNA·RNA hybrids containing purine-rich and pyrimidine-rich strands with DNA and RNA duplexes', *Biochemistry*. doi: 10.1021/bi960948z.

Hafner, M. *et al.* (2010) 'Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP', *Cell*. doi: 10.1016/j.cell.2010.03.009.

El Hage, A. *et al.* (2010) 'Loss of Topoisomerase I leads to R-loop-mediated transcriptional blocks during ribosomal RNA synthesis', *Genes and Development*. doi: 10.1101/gad.573310.

El Hage, A. *et al.* (2014) 'Genome-Wide Distribution of RNA-DNA Hybrids Identifies RNase H Targets in tRNA Genes, Retrotransposons and Mitochondria',

PLoS Genetics. doi: 10.1371/journal.pgen.1004716.

Hall, J. *et al.* (2009) 'Oct4 and LIF/Stat3 Additively Induce Krüppel Factors to Sustain Embryonic Stem Cell Self-Renewal', *Cell Stem Cell*. doi: 10.1016/j.stem.2009.11.003.

Hamperl, S. *et al.* (2017) 'Transcription-Replication Conflict Orientation Modulates R-Loop Levels and Activates Distinct DNA Damage Responses', *Cell*, 170(4), pp. 774-786.e19. doi: 10.1016/j.cell.2017.07.043.

Hamperl, S. and Cimprich, K. A. (2016) 'Conflict Resolution in the Genome: How Transcription and Replication Make It Work', *Cell*. doi: 10.1016/j.cell.2016.09.053.

Hark, A. T. *et al.* (2000) 'CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus', *Nature*. doi: 10.1038/35013106.

Hartono, S. R. *et al.* (2018) 'The Affinity of the S9.6 Antibody for Double-Stranded RNAs Impacts the Accurate Mapping of R-Loops in Fission Yeast', *Journal of Molecular Biology*. doi: 10.1016/j.jmb.2017.12.016.

Harvey, L. *et al.* (2000) *Molecular Cell Biology. 4th edition*, *Journal of the American Society for Mass Spectrometry*. doi: 10.1016/j.jasms.2009.08.001.

Hasegawa, Y. *et al.* (2010) 'The matrix protein hnRNP U is required for chromosomal localization of xist RNA', *Developmental Cell*. doi: 10.1016/j.devcel.2010.08.006.

He, C. *et al.* (2016) ‘High-Resolution Mapping of RNA-Binding Regions in the Nuclear Proteome of Embryonic Stem Cells’, *Molecular Cell*. Cell Press, 64(2), pp. 416–430. doi: 10.1016/J.MOLCEL.2016.09.034.

He, H. *et al.* (2019) ‘Endogenous interaction profiling identifies DDX5 as an oncogenic coactivator of transcription factor Fra-1’, *Oncogene*. doi: 10.1038/s41388-019-0824-4.

He, Y. *et al.* (2021) ‘NF-κB–induced R-loop accumulation and DNA damage select for nucleotide excision repair deficiencies in adult T cell leukemia’, *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.2005568118.

Heger, P. *et al.* (2012) ‘The chromatin insulator CTCF and the emergence of metazoan diversity’, *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.1111941109.

Helbig, R. and Fackelmayer, F. O. (2003) ‘Scaffold attachment factor A (SAF-A) is concentrated in inactive X chromosoe territories through its RGG domain’, *Chromosoma*. doi: 10.1007/s00412-003-0258-0.

Helmrich, A., Ballarino, M. and Tora, L. (2011) ‘Collisions between Replication and Transcription Complexes Cause Common Fragile Site Instability at the Longest Human Genes’, *Molecular Cell*. doi: 10.1016/j.molcel.2011.10.013.

Henras, A. K. *et al.* (2015) ‘An overview of pre-ribosomal RNA processing in

eukaryotes', *Wiley Interdisciplinary Reviews: RNA*. doi: 10.1002/wrna.1269.

Hodroj, D. *et al.* (2017) 'An ATR-dependent function for the Ddx19 RNA helicase in nuclear R-loop metabolism', *The EMBO Journal*, 36(9), pp. 1182–1198. doi: 10.15252/embj.201695131.

Hoeppner, M. A. *et al.* (1996) 'Cloning and characterization of mouse CCAAT binding factor', *Nucleic Acids Research*. doi: 10.1093/nar/24.6.1091.

Hoogsteen, K. (1963) 'The crystal and molecular structure of a hydrogen-bonded complex between 1-methylthymine and 9-methyladenine', *Acta Crystallographica*. doi: 10.1107/s0365110x63002437.

Hou, C. *et al.* (2008) 'CTCF-dependent enhancer-blocking by alternative chromatin loop formation', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.0808506106.

Howe, F. S. *et al.* (2017) 'Is H3K4me3 instructive for transcription activation?', *BioEssays*. doi: 10.1002/bies.201600095.

Hu, E., Liang, P. and Spiegelman, B. M. (1996) 'AdipoQ is a novel adipose-specific gene dysregulated in obesity', *Journal of Biological Chemistry*. doi: 10.1074/jbc.271.18.10697.

Huertas, P. and Aguilera, A. (2003) 'Cotranscriptionally formed DNA:RNA hybrids mediate transcription elongation impairment and transcription-associated recombination', *Molecular Cell*. doi: 10.1016/j.molcel.2003.08.010.

Imbriano, C. *et al.* (2001) 'HSP-CBF is an NF-Y-dependent Coactivator of the Heat Shock Promoters CCAAT Boxes', *Journal of Biological Chemistry*. doi: 10.1074/jbc.M101553200.

Jalal, C., Uhlmann-Schiffler, H. and Stahl, H. (2007) 'Redundant role of DEAD box proteins p68 (Ddx5) and p72/p82 (Ddx17) in ribosome biogenesis and cell proliferation', *Nucleic Acids Research*. doi: 10.1093/nar/gkm058.

Jensen, B. C. *et al.* (2003) 'The NOG1 GTP-binding protein is required for biogenesis of the 60 S ribosomal subunit', *Journal of Biological Chemistry*. doi: 10.1074/jbc.M304198200.

Jeon, Y. and Lee, J. T. (2011) 'YY1 Tethers Xist RNA to the inactive X nucleation center', *Cell*. doi: 10.1016/j.cell.2011.06.026.

Kanehisa, T. *et al.* (1971) 'Studies on low molecular weight RNA of chromatin. Effects on template activity of chick liver chromatin', *BBA Section Nucleic Acids And Protein Synthesis*. doi: 10.1016/0005-2787(71)90511-9.

Kar, B. *et al.* (2011) 'Quantitative nucleolar proteomics reveals nuclear re-organization during stress- induced senescence in mouse fibroblast', *BMC Cell Biology*. doi: 10.1186/1471-2121-12-33.

Kim, H. D., Choe, J. and Seo, Y. S. (1999) 'The sen1+ gene of Schizosaccharomyces pombe, a homologue of budding yeast SEN1, encodes an RNA and DNA helicase', *Biochemistry*. doi: 10.1021/bi991470c.

Klenova, E. M. *et al.* (1993) ‘CTCF, a conserved nuclear factor required for optimal transcriptional activity of the chicken c-myc gene, is an 11-Zn-finger protein differentially expressed in multiple forms.’, *Molecular and Cellular Biology*. doi: 10.1128/mcb.13.12.7612.

Koike, M. *et al.* (2007) ‘Characterization of embryoid bodies of mouse embryonic stem cells formed under various culture conditions and estimation of differentiation status of such bodies’, *Journal of Bioscience and Bioengineering*. doi: 10.1263/jbb.104.294.

Koopman, P. and Cotton, R. G. H. (1984) ‘A factor produced by feeder cells which inhibits embryonal carcinoma cell differentiation. Characterization and partial purification’, *Experimental Cell Research*. doi: 10.1016/0014-4827(84)90683-9.

Lang, K. S. *et al.* (2017) ‘Replication-Transcription Conflicts Generate R-Loops that Orchestrate Bacterial Stress Survival and Pathogenesis’, *Cell*. doi: 10.1016/j.cell.2017.07.044.

Legrand, J. M. D. *et al.* (2019) ‘DDX5 plays essential transcriptional and post-transcriptional roles in the maintenance and function of spermatogonia’, *Nature Communications*. doi: 10.1038/s41467-019-09972-7.

Lehman, B. J. *et al.* (2021) ‘Dynamic regulation of CTCF stability and subnuclear localization in response to stress’, *PLoS Genetics*. doi: 10.1371/journal.pgen.1009277.

Leighton, P. A. *et al.* (1995) ‘An enhancer deletion affects both H19 and Igf2 expression’, *Genes and Development*. doi: 10.1101/gad.9.17.2079.

Lesnik, E. A. and Freier, S. M. (1995) ‘Relative Thermodynamic Stability of DNA, RNA, and DNA:RNA Hybrid Duplexes: Relationship with Base Composition and Structure’, *Biochemistry*. doi: 10.1021/bi00034a013.

Levine, M., Cattoglio, C. and Tjian, R. (2014) ‘Looping back to leap forward: Transcription enters a new era’, *Cell*. doi: 10.1016/j.cell.2014.02.009.

Lex, A. *et al.* (2014) ‘UpSet: Visualization of intersecting sets’, *IEEE Transactions on Visualization and Computer Graphics*. doi: 10.1109/TVCG.2014.2346248.

Li, B. and Dewey, C. N. (2011) ‘RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome’, *BMC Bioinformatics*. doi: 10.1186/1471-2105-12-323.

Li, X. and Manley, J. L. (2005) ‘Inactivation of the SR Protein Splicing Factor ASF/SF2 Results in Genomic Instability’, *Cell*. Cell Press, 122(3), pp. 365–378. doi: 10.1016/J.CELL.2005.06.008.

Li, Y. *et al.* (2020) ‘R-loops coordinate with SOX2 in regulating reprogramming to pluripotency’, *Science Advances*. doi: 10.1126/sciadv.aba0777.

Lieberman-Aiden, E. *et al.* (2009) ‘Comprehensive mapping of long-range interactions reveals folding principles of the human genome’, *Science*. doi: 10.1126/science.1181369.

Linder, P. *et al.* (1989) ‘Birth of the D-E-A-D box [5]’, *Nature*. doi: 10.1038/337121a0.

Linder, P. (2006) ‘Dead-box proteins: A family affair - Active and passive players in RNP-remodeling’, *Nucleic Acids Research*. doi: 10.1093/nar/gkl468.

Liu, K. and Sun, Q. (2021) ‘Intragenic tRNA-promoted R-loops orchestrate transcription interference for plant oxidative stress responses’, *The Plant Cell*. doi: 10.1093/PLCELL/KOAB220.

Lombraña, R. *et al.* (2015) ‘R-loops and initiation of DNA replication in human cells: A missing link?’, *Frontiers in Genetics*. doi: 10.3389/fgene.2015.00158.

Loukinov, D. I. *et al.* (2002) ‘BORIS, a novel male germ-line-specific protein associated with epigenetic reprogramming events, shares the same 11-zinc-finger domain with CTCF, the insulator protein involved in reading imprinting marks in the soma’, *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.092123699.

Lum, L. S. *et al.* (1990) ‘A cloned human CCAAT-box-binding factor stimulates transcription from the human hsp70 promoter.’, *Molecular and Cellular Biology*. doi: 10.1128/mcb.10.12.6709.

Lum, L. S. *et al.* (1992) ‘The hsp70 gene CCAAT-binding factor mediates transcriptional activation by the adenovirus E1a protein’, *Molecular and Cellular Biology*. doi: 10.1128/mcb.12.6.2599-2605.1992.

Luo, H. *et al.* (2020) ‘ Hottip -Mediated R-Loops Regulate CTCF TAD Boundary to Control WNT/b-Catenin Pathway in AML Genome ’, *Blood*. doi: 10.1182/blood-2020-137816.

Lupiáñez, D. G. *et al.* (2015) ‘Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions’, *Cell*. doi: 10.1016/j.cell.2015.04.004.

Lyon, M. F. (1961) ‘Gene action in the X-chromosome of the mouse (mus musculus L.)’, *Nature*. doi: 10.1038/190372a0.

De Magis, A. *et al.* (2019) ‘DNA damage and genome instability by G-quadruplex ligands are mediated by R loops in human cancer cells.’, *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 116(3), pp. 816–825. doi: 10.1073/pnas.1810409116.

Makhlouf, M. *et al.* (2014) ‘A prominent and conserved role for YY1 in Xist transcriptional activation’, *Nature Communications*. doi: 10.1038/ncomms5878.

Mann, M. (2006) ‘Functional and quantitative proteomics using SILAC’, *Nature Reviews Molecular Cell Biology*. doi: 10.1038/nrm2067.

Marino, M. M. *et al.* (2019) ‘Interactome mapping defines BRG1, a component of the SWI/SNF chromatin remodeling complex, as a new partner of the transcriptional regulator CTCF.’, *The Journal of biological chemistry*. American Society for Biochemistry and Molecular Biology, 294(3), pp. 861–873. doi:

10.1074/jbc.RA118.004882.

Marnef, A. and Legube, G. (2020) ‘m6A RNA modification as a new player in R-loop regulation’, *Nature Genetics*. doi: 10.1038/s41588-019-0563-z.

Marnef, A. and Legube, G. (2021) ‘R-loops as Janus-faced modulators of DNA repair’, *Nature Cell Biology*. doi: 10.1038/s41556-021-00663-4.

Martin, G. R. (1981) ‘Isolation of a pluripotent cell line from early mouse embryos cultured in medium conditioned by teratocarcinoma stem cells’, *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.78.12.7634.

Martin, R. *et al.* (2013) ‘DExD/H-box RNA helicases in ribosome biogenesis’, *RNA Biology*. doi: 10.4161/rna.21879.

Mayfield, J. E. and Bonner, J. (1971) ‘Tissue differences in rat chromosomal RNA.’, *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.68.11.2652.

McRae, E. K. S. *et al.* (2017) ‘Human DDX21 binds and unwinds RNA guanine quadruplexes’, *Nucleic Acids Research*. doi: 10.1093/nar/gkx380.

Mersaoui, S. *et al.* (2018) ‘Arginine methylation of DDX5 RGG/RG motif by PRMT5 regulates RNA:DNA resolution’, *bioRxiv*. doi: 10.1101/451823.

Miglietta, G., Russo, M. and Capranico, G. (2020) ‘G-quadruplex-R-loop interactions and the mechanism of anticancer G-quadruplex binders’, *Nucleic*

Acids Research. doi: 10.1093/nar/gkaa944.

Milek, M. *et al.* (2017) 'DDX54 regulates transcriptome dynamics during DNA damage response', *Genome Research*. doi: 10.1101/gr.218438.116.

Moraleva, A. *et al.* (2017) 'Involvement of the specific nucleolar protein SURF6 in regulation of proliferation and ribosome biogenesis in mouse NIH/3T3 fibroblasts', *Cell Cycle*. doi: 10.1080/15384101.2017.1371880.

Morello, L. G., Coltri, P. P., *et al.* (2011) 'The Human Nucleolar Protein FTSJ3 Associates with NIP7 and Functions in Pre-rRNA Processing', *PLoS ONE*. Edited by G. Kudla. Public Library of Science, 6(12), p. e29174. doi: 10.1371/journal.pone.0029174.

Morello, L. G., Hesling, C., *et al.* (2011) 'The NIP7 protein is required for accurate pre-rRNA processing in human cells', *Nucleic Acids Research*. doi: 10.1093/nar/gkq758.

Morgan, A. R. and Wells, R. D. (1968) 'Specificity of the three-stranded complex formation between double-stranded DNA and single-stranded RNA containing repeating nucleotide sequences', *Journal of Molecular Biology*. doi: 10.1016/0022-2836(68)90073-9.

Naftelberg, S. *et al.* (2015) 'Regulation of alternative splicing through coupling with transcription and chromatin structure', *Annual Review of Biochemistry*. doi: 10.1146/annurev-biochem-060614-034242.

Niehrs, C. and Luke, B. (2020) ‘Regulatory R-loops as facilitators of gene expression and genome stability’, *Nature Reviews Molecular Cell Biology*. doi: 10.1038/s41580-019-0206-3.

Nishimura, K. *et al.* (2009) ‘An auxin-based degron system for the rapid depletion of proteins in nonplant cells’, *Nature Methods*. doi: 10.1038/nmeth.1401.

Niwa, H. *et al.* (2009) ‘A parallel circuit of LIF signalling pathways maintains pluripotency of mouse ES cells’, *Nature*. doi: 10.1038/nature08113.

Nora, E. P. *et al.* (2017) ‘Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization’, *Cell*. doi: 10.1016/j.cell.2017.05.004.

Ohle, C. *et al.* (2016) ‘Transient RNA-DNA Hybrids Are Required for Efficient Double-Strand Break Repair’, *Cell*. doi: 10.1016/j.cell.2016.10.001.

Ohlsson, R., Renkawitz, R. and Lobanenkov, V. (2001) ‘CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease’, *Trends in Genetics*. doi: 10.1016/S0168-9525(01)02366-6.

Ong, S. E. *et al.* (2002) ‘Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics.’, *Molecular & cellular proteomics : MCP*. doi: 10.1074/mcp.M200025-MCP200.

Parks, M. M. *et al.* (2018) ‘Variant ribosomal RNA alleles are conserved and exhibit tissue-specific expression’, *Science Advances*. doi:

10.1126/sciadv.aao0665.

Peterson, C. M. *et al.* (2012) ‘Teratomas: A Multimodality Review’, *Current Problems in Diagnostic Radiology*. doi: 10.1067/j.cpradiol.2012.02.001.

Phillips, D. D. *et al.* (2013) ‘The sub-nanomolar binding of DNA-RNA hybrids by the single-chain Fv fragment of antibody S9.6’, *Journal of Molecular Recognition*, 26(8), pp. 376–381. doi: 10.1002/jmr.2284.

Prado, F. and Aguilera, A. (2005) ‘Impairment of replication fork progression mediates RNA polII transcription-associated recombination’, *EMBO Journal*. doi: 10.1038/sj.emboj.7600602.

Pulido-Salgado, M., Vidal-Taboada, J. M. and Saura, J. (2015) ‘C/EBPβ and C/EBPδ transcription factors: Basic biology and roles in the CNS’, *Progress in Neurobiology*. doi: 10.1016/j.pneurobio.2015.06.003.

Rao, S. S. P. *et al.* (2014) ‘A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping’, *Cell*. doi: 10.1016/j.cell.2014.11.021.

Ribeiro de Almeida, C. *et al.* (2018) ‘RNA Helicase DDX1 Converts RNA G-Quadruplex Structures into R-Loops to Promote IgH Class Switch Recombination’, *Molecular Cell*. doi: 10.1016/j.molcel.2018.04.001.

Robbiani, D. F. *et al.* (2009) ‘AID Produces DNA Double-Strand Breaks in Non-Ig Genes and Mature B Cell Lymphomas with Reciprocal Chromosome

Translocations', *Molecular Cell*. doi: 10.1016/j.molcel.2009.11.007.

Robles, J. *et al.* (2005) 'Nucleic Acid Triple Helices: Stability Effects of Nucleobase Modifications', *Current Organic Chemistry*. doi: 10.2174/1385272023373482.

Rocak, S. and Linder, P. (2004) 'Dead-box proteins: The driving forces behind RNA metabolism', *Nature Reviews Molecular Cell Biology*. doi: 10.1038/nrm1335.

Roy, D. and Lieber, M. R. (2009) 'G clustering is important for the initiation of transcription-induced R-loops in vitro, whereas high G density without clustering is sufficient thereafter.', *Molecular and cellular biology*, 29(11), pp. 3124–33. doi: 10.1128/MCB.00139-09.

Sanz, L. A. *et al.* (2016) 'Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals', *Molecular Cell*, 63(1), pp. 167–178. doi: 10.1016/j.molcel.2016.05.032.

Saporita, A. J. *et al.* (2011) 'RNA helicase DDX5 is a p53-independent target of ARF that participates in ribosome biogenesis', *Cancer Research*. doi: 10.1158/0008-5472.CAN-11-1472.

Schmitz, K. M. *et al.* (2010) 'Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes', *Genes and Development*, 24(20), pp. 2264–2269. doi: 10.1101/gad.590910.

Shen, W. *et al.* (2017) ‘Dynamic nucleoplasmic and nucleolar localization of mammalian RNase H1 in response to RNAP I transcriptional R-loops.’, *Nucleic acids research*. Oxford University Press, 45(18), pp. 10672–10692. doi: 10.1093/nar/gkx710.

Shi, J. *et al.* (2013) ‘Role of SWI/SNF in acute leukemia maintenance and enhancer-mediated Myc regulation’, *Genes and Development*. doi: 10.1101/gad.232710.113.

Skene, P. J. and Henikoff, S. (2017) ‘An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites’, *eLife*. doi: 10.7554/eLife.21856.

Skourti-Stathaki, K. *et al.* (2019) ‘R-Loops Enhance Polycomb Repression at a Subset of Developmental Regulator Genes’, *Molecular Cell*. Cell Press. doi: 10.1016/J.MOLCEL.2018.12.016.

Skourti-Stathaki, K., Kamieniarz-Gdula, K. and Proudfoot, N. J. (2014) ‘R-loops induce repressive chromatin marks over mammalian gene terminators’, *Nature*. Nature Publishing Group, 516(7531), pp. 436–439. doi: 10.1038/nature13787.

Skourti-Stathaki, K., Proudfoot, N. J. and Gromak, N. (2011) ‘Human Senataxin Resolves RNA/DNA Hybrids Formed at Transcriptional Pause Sites to Promote Xrn2-Dependent Termination’, *Molecular Cell*. Cell Press, 42(6), pp. 794–805. doi: 10.1016/J.MOLCEL.2011.04.026.

Smolka, J. A. *et al.* (2021) ‘Recognition of RNA by the S9.6 antibody creates

pervasive artifacts when imaging RNA:DNA hybrids', *The Journal of cell biology*. doi: 10.1083/jcb.202004079.

Song, C. *et al.* (2017) 'SIRT7 and the DEAD-box helicase DDX21 cooperate to resolve genomic R loops and safeguard genome stability', *Genes and Development*, 31(13), pp. 1370–1381. doi: 10.1101/gad.300624.117.

Splinter, E. *et al.* (2006) 'CTCF mediates long-range chromatin looping and local histone modification in the β-globin locus', *Genes and Development*. doi: 10.1101/gad.399506.

Srivastava, L. *et al.* (2010) 'Mammalian DEAD Box Protein Ddx51 Acts in 3′ End Maturation of 28S rRNA by Promoting the Release of U8 snoRNA', *Molecular and Cellular Biology*. doi: 10.1128/mcb.00226-10.

Stein, H. and Hausen, P. (1969) 'Enzyme from calf thymus degrading the RNA moiety of DNA-RNA Hybrids: effect on DNA-dependent RNA polymerase.', *Science (New York, N.Y.)*, 166(3903), pp. 393–5. doi: 10.1126/science.166.3903.393.

Stolz, R. *et al.* (2019) 'Interplay between DNA sequence and negative superhelicity drives R-loop structures.', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 116(13), pp. 6260–6269. doi: 10.1073/pnas.1819476116.

Suda, Y. *et al.* (1987) 'Mouse embryonic stem cells exhibit indefinite proliferative

potential', *Journal of Cellular Physiology*. doi: 10.1002/jcp.1041330127.

Szabó, P. E. *et al.* (2000) 'Maternal-specific footprints at putative CTCF sites in the H19 imprinting control region give evidence for insulator function', *Current Biology*. doi: 10.1016/S0960-9822(00)00489-9.

Tang, Y. and Tian, X. (Cindy) (2013) 'JAK-STAT3 and somatic cell reprogramming', *JAK-STAT*. doi: 10.4161/jkst.24935.

Thomas, M., White, R. L. and Davis, R. W. (1976) 'Hybridization of RNA to double stranded DNA: Formation of R loops', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.73.7.2294.

Turner, A. J. *et al.* (2009) 'A Novel Small-Subunit Processome Assembly Intermediate That Contains the U3 snoRNP, Nucleolin, RRP5, and DBP4', *Molecular and Cellular Biology*. doi: 10.1128/mcb.00029-09.

Velichko, A. K. *et al.* (2019) 'Hypoosmotic stress induces R loop formation in nucleoli and ATR/ATM-dependent silencing of nucleolar transcription', *Nucleic Acids Research*. doi: 10.1093/nar/gkz436.

Villarreal, O. D. *et al.* (2020) 'Genome-wide R-loop analysis defines unique roles for DDX5, XRN2, and PRMT5 in DNA/RNA hybrid resolution', *Life science alliance*. doi: 10.26508/lsa.202000762.

Vostrov, A. A., Taheny, M. J. and Quitschke, W. W. (2002) 'A region to the N-

terminal side of the CTCF zinc finger domain is essential for activating transcription from the amyloid precursor protein promoter', *Journal of Biological Chemistry*. doi: 10.1074/jbc.M109748200.

Wahba, L. *et al.* (2016) 'S1-DRIP-seq identifies high expression and polyA tracts as major contributors to R-loop formation.', *Genes & development*. Cold Spring Harbor Laboratory Press, 30(11), pp. 1327–38. doi: 10.1101/gad.280834.116.

Walker, E. *et al.* (2010) 'Polycomb-like 2 Associates with PRC2 and Regulates Transcriptional Networks during Mouse Embryonic Stem Cell Self-Renewal and Differentiation', *Cell Stem Cell*. doi: 10.1016/j.stem.2009.12.014.

Wang, I. X. *et al.* (2018) 'Human proteins that interact with RNA/DNA hybrids'. doi: 10.1101/gr.237362.118.

Wieczorek, S. *et al.* (2017) 'DAPAR & ProStaR: software to perform statistical analyses in quantitative discovery proteomics', *Bioinformatics*. Narnia, 33(1), pp. 135–136. doi: 10.1093/bioinformatics/btw580.

Williams, R. L. *et al.* (1988) 'Myeloid leukaemia inhibitory factor maintains the developmental potential of embryonic stem cells', *Nature*. doi: 10.1038/336684a0.

Wortham, N. C. *et al.* (2009) 'The DEAD-box protein p72 regulates ERα-/oestrogen-dependent transcription and cell growth, and is associated with improved survival in ERα-positive breast cancer', *Oncogene*. doi:

10.1038/onc.2009.261.

Wu, N. Q. and Li, J. J. (2014) ‘PCSK9 gene mutations and low-density lipoprotein cholesterol’, *Clinica Chimica Acta*. doi: 10.1016/j.cca.2014.01.043.

Wutz, A. (2011) ‘Gene silencing in X-chromosome inactivation: Advances in understanding facultative heterochromatin formation’, *Nature Reviews Genetics*. doi: 10.1038/nrg3035.

Xu, W. *et al.* (2017) ‘The R-loop is a common chromatin feature of the Arabidopsis genome’, *Nature Plants*. doi: 10.1038/s41477-017-0004-x.

Yamaguchi, A. and Takanashi, K. (2016) ‘FUS interacts with nuclear matrix-associated protein SAFB1 as well as Matrin3 to regulate splicing and ligand-mediated transcription’, *Scientific Reports*. doi: 10.1038/srep35195.

Yamauchi, T. *et al.* (2014) ‘MDM2 Mediates Nonproteolytic Polyubiquitylation of the DEAD-Box RNA Helicase DDX24’, *Molecular and Cellular Biology*. doi: 10.1128/mcb.00320-14.

Yan, Q. *et al.* (2019) ‘Mapping Native R-Loops Genome-wide Using a Targeted Nuclease Approach’, *Cell Reports*. doi: 10.1016/j.celrep.2019.09.052.

Yang, X. *et al.* (2019) ‘m6A promotes R-loop formation to facilitate transcription termination’, *Cell Research*. doi: 10.1038/s41422-019-0235-7.

Yanling Zhao, D. *et al.* (2016) ‘SMN and symmetric arginine dimethylation of RNA polymerase II C-terminal domain control termination’, *Nature*. doi:

10.1038/nature16469.

Yasuhara, T. *et al.* (2018) ‘Human Rad52 Promotes XPG-Mediated R-loop Processing to Initiate Transcription-Associated Homologous Recombination Repair’, *Cell*. doi: 10.1016/j.cell.2018.08.056.

Yu, K. *et al.* (2003) ‘R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells.’, *Nature immunology*, 4(5), pp. 442–451. doi: 10.1038/ni919.

Yusufzai, T. M. *et al.* (2004) ‘CTCF Tethers an Insulator to Subnuclear Sites, Suggesting Shared Insulator Mechanisms across Species’, *Molecular Cell*. doi: 10.1016/S1097-2765(04)00029-2.

Zaccara, S., Ries, R. J. and Jaffrey, S. R. (2019) ‘Reading, writing and erasing mRNA methylation’, *Nature Reviews Molecular Cell Biology*. doi: 10.1038/s41580-019-0168-5.

Zagulski, M. *et al.* (2003) ‘Mak5p, which is required for the maintenance of the M1 dsRNA virus, is encoded by the yeast ORF YBR142w and is involved in the biogenesis of the 60S subunit of the ribosome’, *Molecular Genetics and Genomics*. doi: 10.1007/s00438-003-0913-4.

Zhang, C., Fu, J. and Zhou, Y. (2019) ‘A review in research progress concerning m6A methylation and immunoregulation’, *Frontiers in Immunology*. doi: 10.3389/fimmu.2019.00922.

Zhou, H. *et al.* (2020) ‘H3K9 Demethylation-Induced R-Loop Accumulation Is Linked to Disorganized Nucleoli’, *Frontiers in Genetics*. doi: 10.3389/fgene.2020.00043.

Zhou, Z., Giles, K. E. and Felsenfeld, G. (2019) ‘DNA-RNA triple helix formation can function as a cis-acting regulatory mechanism at the human β-globin locus’, *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.1900107116.