Characterizing the role of DDX41 in opposing transcription-associated genomic instability

Dissertation

Zur Erlangung des Grades

Doktor der Naturwissenschaften

Am Fachbereich Biologie

Der Johannes Gutenberg-Universität Mainz

Thorsten Mosler

geb. am 03.07.1990 in Frankfurt am Main, Deutschland

Mainz, April 2022

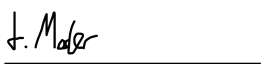
Dekan: Prof. Dr. Eckhard Thines

Tag der mündlichen Prüfung: 01.07.2022

D77

Preface

I, Thorsten Mosler, certify that the presented work and thesis are my own. I confirm that information derived from other sources is indicated throughout the thesis. No help against payment by third parties was accepted.



Thorsten Mosler

Summary

Cycling of cells requires duplication of the genetic information which is encoded in the DNA. During the S-phase of the cell cycle, when DNA replication occurs, cells have to tolerate different types of replication stress. Lesions within the DNA template, DNA secondary structures, or the transcription machinery - through the formation of R-loops - can constitute obstacles for progressing replication forks. Conflicts between transcription-dependent R-loops, which are three-stranded structures consisting of an RNA-DNA hybrid and a displaced non-template DNA strand, and the replication machinery pose a threat to genomic stability. Moreover, R-loops regulate crucial cellular processes such as transcription initiation and termination, chromosome segregation, and DNA repair. Thus, R-loop levels need to be tightly regulated to maintain the regulatory functions as well as genome stability. Loss of R-loop homeostasis due to different cellular perturbations is associated with neurological disorders and cancer. The protein networks that regulate R-loops in human cells are poorly characterized. To this end, we developed RNA-DNA Proximity Proteomics (RDProx) to elucidate the regulatory R-loop proximal proteome in human cells.

Among different nuclear proteins associated with R-loops, we identified the tumor suppressor DEAD-box helicase DDX41. We show that DDX41 can bind and unwind RNA-DNA hybrids *in vitro*. Moreover, DDX41 associates with promoter regions of active genes to maintain R-loop homeostasis and to counteract the accumulation of DNA double-strand breaks *in vivo*. Loss of DDX41 results in R-loop-dependent DNA double-strand breaks, replication stress, and dependency on ATR signaling. R-loop accumulation is accompanied by changes in transcription dynamics and upregulation of inflammatory signature genes.

Germline and somatic mutations in *DDX41* predispose patients to myelodysplastic syndromes and acute myeloid leukemia. Wild type DDX41 opposes DNA damage in CD34+ hematopoietic stem and progenitor cells, while expression of pathogenic variants leads to genomic instability. We propose that loss of promoter-proximal R-loop homeostasis, accumulation of DNA double-strand breaks, and inflammatory signaling contribute to the development of acute myeloid leukemia in individuals with pathogenic DDX41 variants.

Zusammenfassung

Beim Durchlaufen des Zellzyklus muss die genetische Information, die in DNS verschlüsselt ist, verdoppelt werden. Während der DNA Replikation sind Zellen verschiedenen Arten von Replikationsstress ausgesetzt. Neben Schädigungen der DNS-Vorlage und der Bildung von DNS Sekundärstrukturen kann Transkription durch die Formation von R-Loops ein Hindernis für fortschreitende Replikationsgabeln darstellen. Konflikte zwischen transkriptions-abhängigen R-Loops - die aus einem RNS-DNS Hybrid und einem dislozierten DNS Strangs bestehen - und der Replikationsmaschinerie, können zur Instabilität des Genoms führen. Darüber hinaus regulieren R-Loops verschiedene zelluläre Prozesse, wie zum Beispiel Transkription, Chromosomensegregation und DNS-Reparatur. Daher müssen R-Loops streng reguliert werden, um die regulatorischen Funktionen und die Stabilität des Genoms aufrechtzuerhalten. Der Verlust der Homöostase von R-Loops durch verschiedene zelluläre Störungen ist mit neurologischen Störungen und Krebs assoziiert. Proteinnetzwerke, die R-Loops in humanen Zellen regulieren sind unzureichend charakterisiert. Zu diesem Zweck haben wir RNS-DNS Proximal Proteomik (RDPRox) entwickelt, um das regulatorische, proximale Proteom von R-Loops in Humanzellen zu untersuchen.

Unter verschiedenen nuklearen Proteinen, die mit R-Loops assoziiert sind, haben wir die Tumorsuppressor DEAD-box Helikase DDX41 identifiziert, die RNS-DNS Hybride *in vitro* binden und sie entwinden kann. Darüber hinaus assoziiert DDX41 mit hoch transkribierten Promotorenregionen um die Homöostase von R-Loops aufrechtzuerhalten und der Akkumulation von DNS Doppelstrangbrüchen entgegenzuwirken. Der Verlust von DDX41 führt zu einer umfassenden R-Loop-abhängigen Schädigung der DNS, zu Replikationsstress und einer Abhängigkeit von ATR Signalübermittlung. Eine Akkumulation von R-Loops geht mit Änderungen der Transkriptionsdynamik und Hochregulierung von inflammatorischer Signalübermittlung einher.

Durch Keimbahn- und somatische Mutationen in *DDX41* sind Patienten zu myelodysplastischen Syndromen und akuter myeloischer Leukämie prädisponiert. Wildtyp DDX41 wirkt DNS-Schädigungen in CD34+ hämatopoetischen Stamm- und Vorläuferzellen entgegen, während die Expression pathogener genomische Instabilität induziert. Wir schlagen vor, dass der Verlust der Promoter-proximalen R-Loops Homöostase, die Akkumulation von DNS-Doppelstrangbrüchen und die kausale inflammatorische Signalübermittlung zur Entstehung von akuter myeloischer Leukämie in Individuen mit pathogenen *DDX41* Mutationen beiträgt.

Table of contents

1 Introduction

1.1 Cycling cells have to endure various sources of replication stress

DNA replication is a highly regulated process during the S-phase of the cell cycle. Duplication of the genome of eukaryotic somatic cells requires approximately 8-10 hours[1]. This high speed is achieved by facilitating bidirectional replication forks that are initiated from thousands of origins[2]. Replication origins are recognized and marked by the human origin recognition complex (ORC)[3–5]. Coordinated licensing of replication origins is ensured by the minichromosome maintenance (MCM) complex, which is loaded onto the ORC in early G1 to assemble the pre-replicative complex[6,7]. DNA unwinding by the replicative helicase is initiated during the early S-phase through the association of Cdc45 and GINS[8]. Not all licensed origins are fired but kept as dormant backup in case normal replisome progression is impaired[9]. Two bidirectional replication forks converge from each fired origin through the robust activity of the Cdc45-MCM-GINS complex. The Pol a-primase complex initiates DNA synthesis after local unwinding of the helicase complex[10,11]. Subsequently, the leading strand is extended by Pol e, while the lagging strand is synthesized by Pol d[12–16]. DNA strand extension is further supported by factors such as the proliferating cell nuclear antigen (PCNA) and the replication factor C (RFC)[17–19].

DNA replication during the S-phase of the cell cycle poses a threat to genomic stability since it requires local unfolding of the protective chromatin environment. Moreover, loading of the replication machinery leads to increased occupancy of the DNA template. Due to the complexity of DNA replication, cells need to endure several sources of replication stress. Major cell-intrinsic sources of replication stress are obstacles within the DNA template that are encountered by the progressing DNA replication machinery. An aberrant DNA template can be caused by nicks, single-stranded DNA (ssDNA) gaps, mis-incorporated ribonucleotides, or damaged bases (Figure 1). Furthermore, secondary DNA structures such as triplexes caused by trinucleotide repeats or G-quadruplexes (G4s) can alter replication fork progression (Figure 1). Other sources can be the lack of essential replication factors, overexpression of oncogenes, or difficult to replicate regions of the genome such as common fragile sites (Figure 1). Recently, R-loops, which are RNA-DNA hybrids with a displaced ssDNA stretch, have been described as a novel obstacle for the replication machinery. Stabilization of R-loops leads to transcription-replication conflicts (TRCs), which inhibit replication fork progression, thus defining an additional source of replication stress.

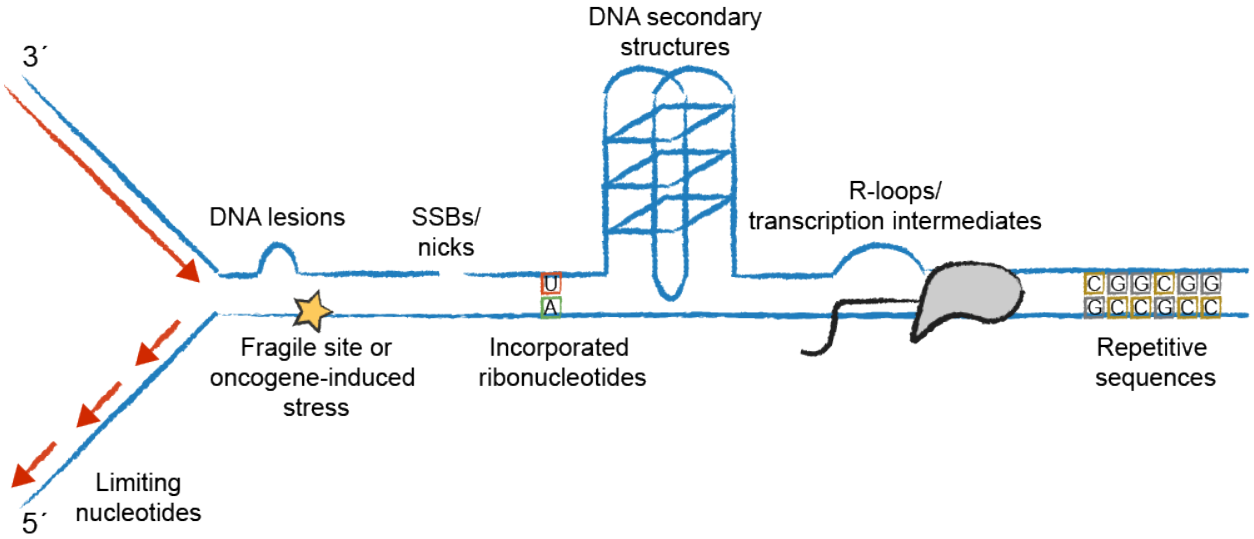


Figure 1: Cycling cells are exposed to various sources of replication stress

Schematics displaying different sources of endogenous replication stress. The DNA double-strand is depicted in blue, nascent DNA as red arrows (the direction of the arrows indicates the direction of DNA synthesis), RNA in black, RNA polymerase in grey, and fragile site as a yellow star. Nucleotides are displayed in different colors: uridine = red, cytosine = yellow, guanine = grey, adenine = red.

Replication fork slow-down or stalling often leads to exposure of ssDNA due to continuous unwinding of the DNA by the replicative helicase, while the DNA polymerases pause[2]. Robust formation of ssDNA stretches induces the recruitment of replication protein A (RPA) which coats and thereby protects exposed ssDNA[20,21]. RPA loading on DNA triggers the recruitment of replication stress response proteins such as the ATR kinase[22]. Recruitment and activation of ATR result in downstream phosphorylation of multiple replication stress response proteins that in return stabilize replication intermediates and promote faithful DNA replication[23]. Prominent substrates of ATR in the replication stress response pathway are the histone variant γH2AX, RPA (Ser33), and the checkpoint kinase 1 (CHK1) (Ser345)[24]. Of note, measurable phosphorylation of these ATR targets requires global replication stress and will not be apparent by only locally stalled replication forks[25]. Globally, ATR promotes cell cycle arrest and suppresses late origin firing, whereas locally ATR stabilizes stalled replication forks to prevent deleterious recombination events[2]. Once the cell manages to remove the replication stress trigger, replication forks can restart and finish DNA synthesis[26]. In case of persistent stalling, cells possess a toolbox to continue replication before resolving the replication stress source[2]. Replication forks can be stabilized by fork reversal[27], replication can be continued by firing dormant origins[28] or by repriming downstream of a lesion[29]. Moreover, DNA lesions can be bypassed by the replisome using translesion synthesis[30] or template

switching[31]. Unfaithful stabilization of stalled replication forks results in replication fork collapse, caused by dissociation or run-off of the replication machinery. Furthermore, reversed forks can be a target of nuclease-dependent degradation[32].

1.2 RNA Transcription is a highly orchestrated process

In human cells, DNA-dependent RNA transcription is carried out by three distinct RNA polymerases[33]. RNA Pol I is responsible for the transcription of pre-rRNA, whereas RNA pol III transcribes transfer-RNAs (tRNAs) and other non-coding RNAs[34–36]. Pre-mRNA-coding genes are transcribed by RNA polymerase II, which consists of 12 subunits[37–39]. RNA synthesis is initiated by the assembly of the pre-initiation complex (PIC) at promoter regions[40]. The PIC assembles through the association of RNA Pol II and various transcription factors such as TFIIB and TFIIH, which promote RNA Pol II release from the promoter and motion towards the gene body[39,41–44]. Binding of transcription factors to DNA is often sequence-specific (e.g. TATA box)[45–47]. Signals from DNA-associated transcription factors to the RNA polymerase II are propagated via the 26-subunit mediator complex[44,48]. Moreover, the mediator complex contributes to the organization of topological DNA structures such as chromatin loops, which are required for enhancer function[49,50]. After transcription initiation, RNA Poll II pauses 30-60 base pairs (bps) downstream of the transcription start site (TSS)[51–53]. RNA Pol II pausing is required to ensure the integrity of the 5′-cap of the nascent transcript and is crucial to prepare RNA Pol II for faithful transcription elongation. Promoter-proximal pausing is mediated by negative elongation factors (NELFs), which restrict Pol II to the pause site[54,55]. Subsequently, CDK9 and positive transcription elongation factor-b (pTEFb) induce the release of paused RNA Pol II, thereby promoting transcription elongation throughout the gene body[56–59]. Different phosphorylation states of the C-terminal domain (CTD) of RNA Pol II mark the different stages of the transcriptional cycle[51]. During promoter-proximal pausing the CTD is hyper-phosphorylated on serine 5 residue[60,61]. De-phosphorylation of Ser5 is accompanied by increased phosphorylation on serine 2, which is the dominant phosphorylation state during productive elongation[51,62,63]. The speed of RNA Pol II during transcription elongation can be modified to regulate co-transcriptional processes such as splicing[54,64–67]. Vice-versa, aberrant splicing has been linked to reduced elongation speed and stalling of RNA Pol II[68,69]. Once RNA polymerase elongation is complete, transcription is terminated within specific regions. Transcription termination sites are distinct regions behind the coding region that serve as boundaries to define genes as modular elements and display an

important regulatory step during gene expression. The choice of termination pathway is a key determinant for the faith of mature transcripts[70]. Termination of mRNA transcription usually triggers nuclear export and subsequent translation, while termination of non-coding RNAs (ncRNAs) often causes nuclear retention or degradation[70]. Moreover, early termination sites within genes result in shorter transcript isoforms with distinct functions from the full-length transcript[71]. Impaired release of the polymerase from the promoter-proximal pause site, stalling during the elongation process, or unfaithful transcription termination are associated with the formation of R-loops and transcription-dependent genomic instability[72–74].

1.3 Discovery of R-loops

Secondary loop structures within DNA molecules have already been described five decades ago[75]. So-called displacement loops (D-loops) were discovered as intermediates during replication of mitochondrial DNA (mtDNA)[75,76]. Intriguingly, D-loops are not randomly generated but instead form at distinct sites of the circular mtDNA. A regulatory function of D-loops during priming of mtDNA replication was subsequently identified. In 1976, Thomas et al. observed that a complementary RNA molecule can hybridize to a double-stranded DNA (dsDNA) molecule[77]. Negative supercoiling of the DNA allows the RNA to invade the DNA duplex, thereby forming an RNA-DNA hybrid and at the same time displacing the non-template DNA strand. Due to their similarity to the previously reported D-loops and their RNA moiety, these non-B-DNA structures were named R-loops. After studies of RNA-DNA hybrid formation during the lifecycle of adenoviral RNA, the first monoclonal antibody for immunodetection of RNA-DNA hybrids was characterized[78–80]. The first proof that R-loops not only form artificially during *in vitro* reactions but also *in vivo* was provided in 1995. Drolet et al. reported that R-loops form during transcription in *E. coli* and that mutations in the *topA* gene lead to significant growth defects[81]. Moreover, overexpression of RNase H, which removes RNA-DNA hybrids, partially rescued the phenotype, suggesting that the decrease in bacterial fitness was caused by deleterious R-loop formation. Providing evidence for R-loops *in vivo* and discovering the first R-loop regulators in bacteria set the foundation for a new research field intending to understand the molecular function of R-loops and their regulation.

1.4 R-loops are defined by specific properties

R-loops are three-stranded nucleic acid structures consisting of an RNA-DNA hybrid and a displaced non-template ssDNA moiety. The RNA-DNA hybrid core is neither forced into the A-

form of double-stranded RNA (dsRNA) nor the B-form of dsDNA[82,83]. Instead, RNA-DNA hybrids form an intermediate state between the two forms. Once annealed, RNA-DNA hybrids are thermodynamically more stable than dsDNA *in vitro*[82,84,85]. Different R-loop lengths on chromatin have been described based on the method of identification and the location in the genome. Electron microscopy (EM) revealed that R-loops span a length between 150-500 bps[86]. A median R-loop length of approximately 300 bps was determined by single-molecule R-loop footprinting[87]. Longer hybrids of a median of 347 bp have been observed in the promoter region and shorter hybrids of a median of 329 bp in gene bodies[87]. In extreme cases, R-loops can span up to 2.7 kb[87]. The formation of R-loops is conserved across various species, ranging from bacteria to yeast and mammals[88]. R-loops are quite abundant on human chromatin. Approximately 150 mega bases (Mb) or 5% of the human genome are occupied by R-loops[89]. Certain genomic features have been discovered to promote R-loop formation. Sequences with high GC-skew favor R-loop formation by stabilizing the displaced ssDNA via G4 folding[90]. Accordingly, treatment with G4 stabilizing ligands leads to R-loop accumulation and consequent DNA damage[90,91]. Other genomic features that impact R-loop formation are the transcriptional status of a gene and repetitive sequences, such as centromeres, telomeres, and enhancers[73,92–95]. Additionally, high occupancy of R-loops at promoters and termination sites has been revealed by genome-wide sequencing approaches[89,96]. DNA lesions also promote R-loop formation. Single strand breaks in the DNA can serve as R-loop initiation zones, independently of the G-richness of the sequence[97]. Recently, various studies suggested that RNA-DNA hybrids form in the vicinity of double-strand breaks (DSBs)[98–101].

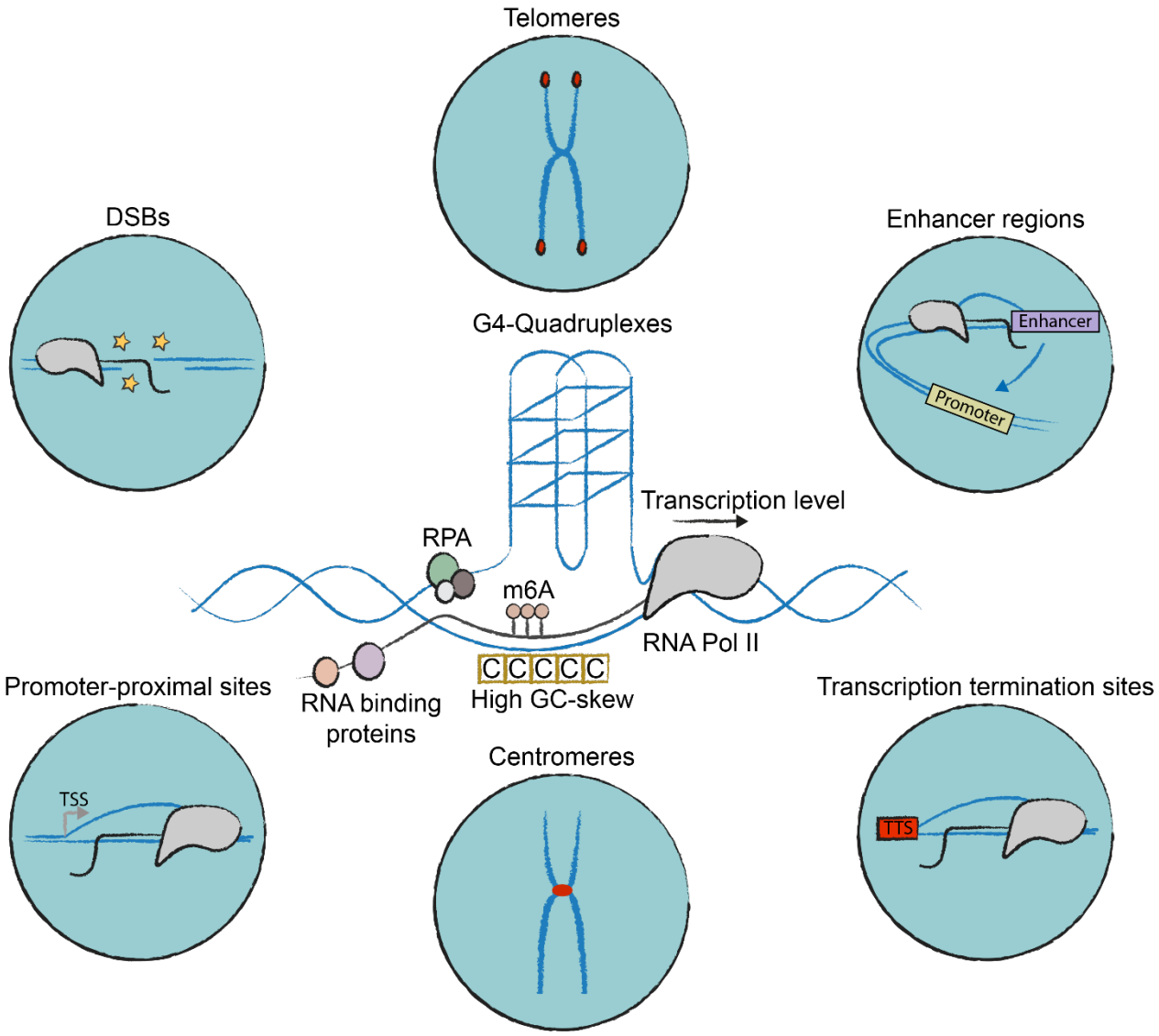


Figure 2: R-loops form at distinct loci within the genome and are stabilized by different mechanisms

Schematics displaying features that promote R-loop stabilization (indicated in the center) and genomic regions that are prone to R-loop formation (depicted in turquoise circles). DNA is displayed in blue, telomeres and centromeres as red dots, RNA polymerase in grey, RNA in black, transcription start site (TSS) as grey arrow, transcription termination site (TTS) as red box, and cytosines in yellow boxes.

1.5 The formation of R-loops is highly associated with transcriptional activity

Different mechanisms describing how R-loops form have been proposed. Most R-loops *in vivo* form co-transcriptionally[73,102,103]. Negative supercoiling of the DNA is generated behind RNA Pol II during transcription, allowing the newly synthesized transcript to invade the DNA double-strand. Hybridization of the RNA to the DNA displaces the non-template DNA strand and leaves it behind as a single strand. Not only transcription by RNA Pol II can result in R-loop formation. Activity of RNA Pol I at highly-transcribed ribosomal DNA (rDNA) loci results in the formation of R-loops[104].

Accordingly, loss of yeast Top1 results in transcription-blockage due to the accumulation of inherent R-loops formed by ribosomal RNA (rRNA) during pre-rRNA synthesis[104]. Transcriptional activity of RNA Pol III is also associated with R-loop formation. During transcription of tRNA in yeast, RNA Pol III-dependent R-loops are stabilized in the absence of the major R-loop regulators rnh1, rnh201 and sen1-1[105]. Furthermore, sen1 is crucial for transcription termination of RNA Pol III, suggesting regulatory R-loop formation during tRNA synthesis[106]. Transcription-dependent R-loops are thought to form stochastically *in cis*. However, also mechanisms that actively promote R-loop formation *in trans* have been proposed[107]. R-loop formation *in trans* occurs during bacterial adaptive immunity and genome-editing activity of clustered, regularly interspaced, short palindromic repeats (CRISPR)–associated protein Cas9. A guide RNA (gRNA) within the CRISPR-CAS9 complex initiates a short homology search that, if successful, drives local unwinding of the target DNA double strand within the Cas9 protein[108,109]. Subsequent strand invasion of the gRNA into the dsDNA causes formation of an *in trans*-derived R-loop within the ribonucleoprotein (RNP) complex. The displaced ssDNA moiety is thereby positioned close to the RuvC domain. These events in return cause NHN nuclease domain to locate to the DNA strand cleavage site, which is the crucial conformation for faithful DNA cutting[108]. This suggests that *in trans*-generated R-loops depend on an active homology search and strand invasion mechanism facilitated by dedicated protein machinery. A class of proteins that can overcome these mechanistic challenges is the homologous recombination (HR) machinery. In budding yeast, mutants that are defective for transcription repression and RNA degradation, Rad51p and Rad52p are crucial for the formation of RNA-DNA hybrids[110]. Furthermore, Rad51p mediates R-loop formation of RNA species distal from their locus of synthesis, indicating transcription-independent hybridization *in trans*. Importantly, the deleterious R-loop formation is suppressed by the Rad51p antagonist Srs2p[110]. Similarly, a potential gene regulatory mechanism for long non-coding RNAs (lncRNAs) based on R-loops formed *in trans* has been described[111]. The formation of *Gal* lncRNA-DNA hybrids was linked to environmental adaptation and transcriptional regulation. The RNA helicase Dbp2 promotes the transcription-independent formation of lncRNA-DNA hybrids across the *Gal* gene cluster, providing evidence for an additional class of R-loops formed *in trans*[111]. The quick adaptation through *Gal* lncRNA transcriptional regulation is suggested to provide a competitive fitness advantage over carbon source shift[111]. In addition to the clear mechanistic separation between *in cis* and *in trans*-induced R-loops, a *cis/trans* model has been proposed[107]. As part of this mixed model, a co-transcriptional R-loop formed *in cis* causes displacement of the non-

template ssDNA, which in return could be used by a complementary RNA for *in trans* hybridization[107]. This mechanism would simply require a protein with strand-annealing ability, such as Rad52 since strand invasion becomes unnecessary due to the exposure of the ssDNA moiety.

1.6 Regulatory roles of R-loops

1.6.1 R-loops promote faithful chromosome segregation during mitosis

Repetitive sequences within the genome are prone to R-loop formation. Centromeric repeats are repetitive elements within the human genome that are highly occupied by R-loops[95,112]. A conserved feature of centromeres from yeast to humans is the cell cycle-dependent transcription of non-coding RNAs that enable R-loop formation[73,95,113]. During mitosis, the ssDNA patch of centromeric R-loops is covered with the ssDNA binding protein RPA[95]. CENP-F recruits ATR to centromeric R-loops in an Aurora kinase A (AURKA)-dependent manner[95]. RPA bound to the displaced ssDNA moiety stimulates the activity of recruited ATR, which subsequently activates Aurora kinase B (AURKB) via the CHEK1 kinase[95]. This R-loop-driven function of ATR at centromeres promotes faithful chromosome segregation independently of its role in opposing replication stress and DNA damage[95]. Furthermore, the mitotic R-loop landscape differs significantly from the interphase landscape. R-loops dynamically accumulate on chromosome arms during prophase and are quickly resolved in an AURKB-dependent manner before centromeric R-loops form during a mitotic stall[112]. The RNA-binding protein RBMX subsequently localizes AURKB to the centromeres, thereby promoting the recruitment of the chromosome passenger complex (CPC) to the inner centromere. Proper localization of the CPC is promoted by RBMX in an R-loop-dependent manner and ensures faithful mitosis. These findings highlight the importance of centromeric R-loop formation during mitosis and emphasize the importance of R-loop promoting repetitive elements for maintaining proliferation.

1.6.2 Class-switch recombination is dependent on R-loop formation

Class-switch recombination (CSR) is a site-directed gene rearrangement mechanism that is crucial for proper immune function in vertebrates[114,115]. Recognition of an antigen activates CSR in B cells to produce different species of antibodies as part of the immune response. Hotspots for B cells with active CSR are the lymph nodes and the gut-associated lymphoid system[114]. During CSR, the isotype of immunoglobulin can be switched from the initial μ exon to various other exons such as α, y, or ε, leading to an exchange of the heavy chain constant region, thereby influencing the

functional properties of the mature antibodies[116]. Due to the flexible switching mechanism, IgH, IgA, or IgG can be produced, while keeping the antigen specificity constant. Therefore, CSR provides an efficient tool to react to certain antigens with differentiated antibody responses.

Immunoglobulin switch regions are highly transcribed right before the recombination sites[117]. *In vitro* transcription of a plasmid containing the murine *IgA* switch locus leads to the loss of super helical turns of the DNA and results in excessive formation of RNA-DNA hybrids through nascent RNA transcripts[118]. These findings suggest a regulatory role of R-loops during CSR. Further investigations confirmed that R-loop formation in highly repetitive immunoglobulin switch regions is a major driver for the recombination events that are required to initiate the switch of the immunoglobulin isotypes *in vivo* [114,119]. Stimulated B cells express long R-loops with a displaced Guanine-rich ssDNA strand in actively transcribed switch regions[114]. R-loop formation is promoted by the repetitive sequence in addition to the high GC-skew within the 1-12 kilo base (kb) switch regions[120]. One of the crucial enzymes during CSR is the activation-induced cytidine deaminase (AID)[121]. Overexpression of AID in B lymphoma cells leads to initiation of CSR and spontaneous switching from IgM to IgA without cytokine induction[122]. Additionally, patients with *AID* mutations suffer from the autosomal recessive form of hyper IgM syndrome, caused by alterations during class switching to other heavy chain isotypes[123]. Moreover, AID is recruited to CSR loci by G4-forming ncRNA species that act as guides after lariat formation[124]. The DEAD-box helicase DDX1 is essential for RNA-mediated AID recruitment in B cells. DDX1 resolves the G4 structure within the AID-bound ncRNAs and thereby promotes R-loop formation *in trans* at CSR loci and contributes to recruiting AID[119]. After recruitment to DNA, AID acts on the ssDNA patch of class switch region R-loops and converts dC to dU[125]. As a consequence, a DSB is generated and repaired by non-homologous end joining (NHEJ), thereby completing CSR[121]. To this end, it is not entirely understood how the DSB is induced. Santos-Peirrera and colleagues proposed a mechanism during which the deamination of dC to dU by AID attracts the base excision repair or the mismatch repair machinery, which processes these sites into nicks, which ultimately are converted to DSBs[121]. R-loop formation in CSR regions constitutes a prominent example of regulatory functions of R-loops within genomic repetitive elements.

1.6.3 Telomere maintenance is promoted by R-loop formation

Another class of repetitive elements in the genome that promotes R-loop formation are telomeres, which are located at the end of linear chromosomes[92,126]. Telomeric TTAGGG repeats are crucial

to maintaining genome stability by protecting chromosome ends from being recognized by the DNA repair machinery through the recruitment of protective proteins such as the shelterin complex[127]. If not properly protected, telomeres can be recognized as DSBs and subsequently be recombined with other accessible chromosome ends. Subsequent fusion of chromosome ends is a major driver of genomic instability in cells with insufficient telomere protection[128,129]. In proliferating cells, telomeres shorten over time due to the end replication problem[130]. Once they reach a critically short state, healthy somatic cells will induce senescence to prevent tumorigenic recombination events[131,132]. In contrast to somatic cells, germ cells need to be preserved, and senescence needs to be prevented. Thus, germ cells elongate their telomeres to circumvent the end replication problem, thereby escaping senescence. Elongation of telomeres is ensured by an RNP named telomerase[133–135]. Due to their high replication rates, cancer cells run more quickly into the end replication problem. They hijack the function of telomerase to prevent the onset of senescence, thereby gaining immortality[136,137]. Roughly 10%-15% of cancer cells lack functional telomerase, while still being able to elongate their telomeres[138]. This sub-class of cancers immortalizes through a mechanism called alternative lengthening of telomeres (ALT)[139]. Not only cancer cells maintain telomere length by facilitating ALT. In budding yeast, ALT has been reported in the absence of a functional telomerase[140]. If telomeres get critically short, a subset of cells overcomes senescence by performing ALT. This mechanism is R-loop-dependent since overexpression of RNase H leads to a quicker onset of senescence[73,94]. Telomeric R-loops are formed by the telomere repeat-containing RNA (TERRA)[141]. The G-rich sequence of TERRA is most likely the driver of R-loop formation due to stabilization by G4 formation on the displaced ssDNA strand[140]. TERRA R-loops have been proposed to be transiently formed and display cell cycle-dependent characteristics. Hybrids formed by TERRA are dynamically regulated during S-phase, while being removed by RNase H during late S-phase when telomeres need to be replicated[92,141]. When telomeres get critically short over time, TERRA R-loops are stabilized. These less transient R-loops trigger recombination of a short telomere with a longer telomere, thereby preventing chromosome end fusions and premature onset of senescence[73]. Fine-tuning of R-loop levels in ALT-positive cancer cells is crucial for their survival. The major R-loop regulator RNaseH1 controls telomeric hybrid levels in a moderate fashion in order to maintain constitutive telomeric replication stress. Downstream loading of the ssDNA-binding protein RPA triggers telomeric HR with longer telomeres[142]. Loss of RNaseH1 leads to elevated R-loop levels at telomeres, which cause excessive telomeric replication stress and RPA loading. Deleterious levels of telomeric replication stress in

the absence of RNaseH1 ultimately lead to telomere loss[142]. When RNaseH1 is overexpressed, telomeric R-loop formation is prevented and telomeres get critically short over time in telomerase-negative cancer cells[142]. Additionally, the ATPase/translocase FANCM has been shown to balance telomeric replication stress in ALT cells[143]. Loss of FANCM results in dysfunctional telomeres and consequent cell death. These findings highlight the importance of R-loops and their strict regulation in the maintenance of telomeres in telomerase-negative cancer cells.

1.6.4 R-loops contribute to transcription termination

An important step during the transcriptional cycle of RNA polymerases is the termination of transcription. During termination, the newly synthesized RNA transcript is uncoupled from the RNA polymerase, which itself is released from the DNA template[144]. Successful transcription termination is crucial for faithful post-transcriptional pre-mRNA processing and gene expression[145]. Aberrant termination leads to read-through transcription and the occurrence of "downstream of genes" (DOGs) transcripts. Since DOGs can extend until reaching neighboring genes, they are thought to be deleterious through interference effects[146]. In order to terminate transcription successfully, the transcription machinery relies on a poly(A) signal and a downstream terminator sequence[144,145,147,148]. Two major classes of terminator sequences exist in human cells. One class relies on XRN2 activity to cleave the RNA transcript co-transcriptionally to release it[148]. The second class of termination elements requires a G-rich sequence, which establishes a pause site for the RNA polymerase downstream of the poly(A) signal that is crucial for the efficient release of the transcript and the polymerase[149]. Genome-wide GC-skew analysis of the human genome revealed that the termination sites downstream of the 3´-end of roughly 2000 genes contained GC-rich sequences[150]. Genes with high GC-skew termination sites are enriched in areas of the genome with high gene density, in which faithful transcription termination is crucial to prevent interference of read-through transcription with neighboring genes. Sequences containing high GC-skew are prone to R-loop formation, suggesting a potential role of R-loops at G-rich transcription termination sites. Indeed, genome-wide R-loop sequencing approaches revealed enrichment of R-loops at transcription termination sites[87,151]. R-loops within termination sequences are more than co-transcriptional byproducts, but fulfill regulatory roles during transcription termination. The formation of R-loops downstream of the poly(A) signal promotes stalling of RNA Pol II, thereby establishing a pause site within the terminator region[73,152]. Pausing of RNA Pol II is crucial to efficiently terminate transcription at G-rich sequences downstream of the poly(A) signal.

R-loop-dependent tethering of RNA Pol II to chromatin allows the 5´-3´exoribonuclease XRN2 to cleave the nascent transcript to release it from the polymerase[121,153]. A second termination factor is the RNA/DNA helicase senataxin (SETX) which modulates R-loop levels within the transcription termination region[144]. Excessive R-loop stabilization downstream of the poly(A) signal in the absence of SETX leads to a significant amount of read-through transcription[144]. As a consequence, gene expression is highly altered due to interference with the transcription of neighboring genes. Fine balancing of R-loops within terminator regions is crucial to ensure faithful transcription termination and release of the nascent transcript. Low levels of R-loops fail to efficiently pause the RNA polymerase, whereas excessive R-loop formation results in sustained tethering of RNA polymerases to chromatin. Similar to SETX, the DEAH-box helicase DHX9 has been reported to suppress R-loop accumulation in transcription termination regions, thereby protecting the genome against read-through transcription of RNA Pol II[152]. The mechanism of how R-loops cause RNA Pol II to pause downstream of the 3´-end of genes is poorly understood. One hypothesis is that an overextended RNA-DNA hybrid downstream of the poly(A) sequence leads to the de-stabilization of the RNA Pol II elongating complex, thereby terminating transcription due to conformational changes of the DNA template[73,154]. An alternative hypothesis is based on alternating RNA Pol II between the elongating and the backtracking state. The formation of RNA-DNA hybrids at G-rich termination sites constitutes an obstacle for backtracked RNA Pol II and thus prevents switching back to the elongating state, thereby pausing the polymerase[73]. In addition, another model suggests that RNA Pol II pausing in GC-rich termination regions is caused by changes in the chromatin architecture[155]. R-loop formation downstream of the poly(A) signal triggers antisense-transcription over the pause elements. Newly synthesized antisense transcripts form double-stranded RNAs (dsRNAs), which in return recruit RNA processing factors such as DICER and AGO1/2, in addition to the G9a histone lysine methyl transferase. G9a establishes a repressive H3K9me2 mark, thereby attracting hetero-chromatin protein 1y (HP1y), which induces heterochromatin formation and consequent RNA Pol II pausing[155].

Diverse mechanisms of how R-loop formation promotes transcription termination of RNA Pol II have been described. Until today, it is not entirely clear how R-loops establish a pause site at GC-rich termination sequences. However, it can be stated that R-loop levels downstream of the poly(A) signal need to be tightly regulated since loss of R-loops and excessive formation lead to read-through transcription and subsequent interference with neighboring genes.

1.6.5 R-loops promote transcription initiation

R-loops do not only form downstream of the 3′-end of human genes but also proximal to the 5′-end. Genome-wide R-loop mapping revealed that R-loops form excessively around the transcription start site (TSS) and the promoter region[87,96,156]. Especially CpG island promoters (CGIs) are prone to R-loop formation due to their high positive GC-skew on the non-template strand[157]. As in other parts of the genome, R-loops do not only form around the TSS as transcriptional by-products but carry out specific regulatory functions to modulate transcriptional activity. DNA cytosine methylation within CGIs is a well-established mechanism of cells to silence specific genes to establish certain gene expression patterns[158,159]. Repressive methylation marks are used to achieve tissue-specific expression and allow the inheritance of gene expression regulation. Genome-wide analysis revealed that 80-90% of CGI promoters are un-methylated in healthy tissue and dysregulation of methylation is implicated in cancer development[160]. Suppression of transcription by hyper-methylation of CGIs of classical tumor suppressor genes such as *BRCA1* contributes to the development of HR-deficient triple-negative breast cancers[161]. Similarly, altered *MLH1* expression by increased methylation of its gene promoter has been proposed as a driver of colorectal carcinoma[162]. These examples highlight the importance of proper maintenance of the CGI methylation status to maintain genome integrity. R-loop formation plays an important role in preventing methylation of CGIs[157]. The RNA-DNA hybrid moiety within R-loops reduces the binding affinity of DNA methyltransferases (DNTMs) to nucleic acids i*n vitro*[157]. Accordingly, CGIs that are occupied by R-loops repel DNTMs and thus prevent *de-novo* methylation and gene silencing[157]. Cells recovered from amyotrophic lateral sclerosis 4 (ALS4) patients carrying a somatic L389S mutation within *SETX* display globally reduced R-loop levels[163]. As a consequence of decreased R-loop formation, over 1200 genes show increased promoter methylation in cells expressing the pathogenic L389S SETX variant. The hyper-methylation is a result of increased DNMT1 recruitment to gene promoters when R-loops are absent since DNMT1 possesses a higher binding affinity towards dsDNA than RNA-DNA hybrids[163]. In addition to reducing the association of DNTMs to CGIs, R-loops actively promote DNA de-methylation by recruiting the responsible machinery. Antisense R-loop formation by the lncRNA *TARID* at the *TCF21* promoter attracts the growth arrest and DNA damage protein 45A (GADD45A), which subsequently promotes DNA de-methylation by recruiting the TET1 enzyme[164]. The binding of GADD45A to the antisense R-loop causes loss of CGI methylation and thereby induces sense *TCF21* transcription. Genome-wide profiling in embryonic stem cells further revealed thousands of R-loop-dependent TET1 binding

sites within CGIs, suggesting a role of R-loops in epigenetic regulation by promoting active DNA de-methylation in a GADD45A-dependent manner[164]. Activation of transcription by antisense ncRNAs has also been described for the *VIM* gene in colon tumors[165]. Expression levels of an antisense transcript are positively correlated with a head-to-head sense transcript[165]. Hyper-methylation of the *VIM* promoter is dependent on antisense R-loop formation and results in higher transcription rates of the sense transcript. Repression of R-loop formation by antisense knockdown or R-loop destabilization by RNaseH1 overexpression leads to chromatin compaction and reduced binding of the transcriptional activators of the NF-kB pathway[165]. These findings underline the role of R-loop formation in regulating the transcriptional activity through modification of the methylation status of gene promoters, either by *blocking de-novo* methylation or by actively promoting promoter de-methylation.

1.6.6 Double-strand break-associated R-loops promote DNA repair

R-loops have been observed in the vicinity of DSBs within actively transcribed regions of the genome[99,100,166,167]. However, it is not entirely understood whether R-loop formation at DSBs is actively promoted by cells, or if they form as a consequence of transcriptional stalling at DNA lesions[168]. Recruitment of catalytically-dead RNaseH1 to micro-laser irradiation tracts provides further evidence for R-loop formation at DSBs[168,169]. Moreover, DSB-associated R-loops are also induced by other sources than irradiation. Cells treated with the DSB-inducing drug zeocin display elevated R-loop formation around the break site[170]. Furthermore, reactive oxygen species (ROS) cause the formation of RNA-DNA hybrids in actively transcribed regions of the genome, triggering a CSB-dependent homologous recombination pathway that is distinct from the canonical BRCA1/2-dependent HR pathway[171]. These findings suggest that R-loops form at DSBs independent of the source of damage. However, the reasons why R-loops form at DSBs remain elusive. A potential explanation for R-loop occurrence proximal to DSBs is the pausing of RNA Pol II caused by transcription pausing at lesions within the DNA template[73,101,168,172]. In line, reduced phosphorylation of RNA Pol II CTD-Ser2 around nuclease-induced DSBs is evident, which indicates reduced transcriptional elongation[173]. Decreased phosphorylation of CTD-Ser2 is accompanied by increased phosphorylation of CTD-Tyr1, which has been reported to drive transcription termination[168,174–176]. These changes in the CTD phosphorylation pattern indicate reduced transcriptional activity in the vicinity of DSBs in RNA Pol II-transcribed genomic regions. In addition, the negative elongation factor NELF is recruited to DSBs, thereby contributing to

increased RNA Pol II pausing[177]. Moreover, a repressive chromatin state is established around DSBs, thus promoting transcriptional shutdown[178–180]. Taken together, these findings suggest that DSBs within actively RNA Pol II-transcribed regions lead to stalling of elongating RNA Pol II. Reduced elongation and consequent R-loop formation constitute a potential mechanism for the stabilization of DSB-proximal RNA-DNA hybrids. In contrast, multiple studies propose active recruitment of the transcription machinery to generate RNA-DNA hybrids at DSBs as part of the DNA damage response[100,166,168,170]. One possible model of damage-induced RNA-DNA hybrids is the formation *in cis* during which DSB ends serve as non-canonical promoters[166]. In *Schizosaccharomyces pombe*, RNA Pol II and RNA-DNA hybrids accumulate around a site-specifically-induced DSB[166]. After end resection by the MRN complex, RNA Pol II utilizes the 3´overhang of the DNA template to initiate transcription, thereby inducing the formation of RNA-DNA hybrids around the break site. On the one hand, active transcription and hybrid formation are crucial to prevent excessive resection and loss of genetic information in the vicinity of the DSB, on the other hand, the RNA moiety needs to be efficiently removed to allow RPA loading and subsequent DSB repair. Thus, tight regulation of DSB-induced RNA-DNA hybrid levels at break sites is crucial for cells[166]. In human U2OS and Hela cells, induction of a nuclease-induced DSB causes recruitment of a complete RNA Pol II initiation complex, including CDK7 and CDK9[167]. The resection-dependent DNA 3´-overhang at DSBs establishes a functional sequence-independent promoter that triggers transcription of damage-induced long non-coding RNAs (dilncRNAs)[167]. DNA repair foci formation is dependent on dilncRNAs since loss of the pre-initiation complex and subsequent loss of dilncRNAs disrupts 53BP1 and pATM foci in response to DNA damage[167]. The authors propose that DSB-induced transcriptional promoters and actively-transcribed dilncRNAs serve as a seed to mediate phase separation of DNA damage repair factors into foci[167]. The hypothesis that RNA Pol II is actively recruited to DSBs is further supported by the finding that the MRN complex and RNA Pol II are sufficient to initiate the synthesis of non-coding RNA species in an *in vitro* reconstituted system[181]. Loading of RNA Pol II and transcription at the DNA end is dependent on the DNA melting activity of the MRN complex and is further promoted when RPA is added to the reaction[181]. Stabilization of the ssDNA by RPA after successful end resection by MRN allows RNA Pol II to associate more efficiently. Damage-induced RNA-DNA hybrids formed by dilncRNAs have also been reported to contribute to the HR pathway during the S/G2 phase of the cell cycle[182]. After end resection of the DSB dilncRNAs associate with the DNA 3´-overhang and participate in attracting BRCA1 to the break site. Subsequently, other HR factors

such as BRCA2 and RAD51, are recruited. The residual RNA-DNA hybrid is later removed by RNaseH2 in a BRCA2-dependent manner. Moreover, SETX and DDX5 helicases are recruited to DSBs to remove RNA-DNA hybrids to allow faithful DNA repair[101,183]. Taken together, these reports indicate that resected DSBs can serve as sequence-independent promoters that lead to the transcription of dilncRNAs which form RNA-DNA hybrids *in cis*. Recruitment of CDK7 and CDK9 suggests that damage-induced transcription follows the canonical mode[168]. Damage-induced RNA-DNA hybrids in the vicinity of DSBs participate in various repair mechanisms and need to be removed to allow faithful DSB repair.

1.6.7 R-loops regulate DNA replication initiation

R-loops have been described to initiate replication in various contexts. The first evidence that RNA-DNA hybrids promote the onset of replication was provided in the 1980s. Daspugta and colleagues reported that an RNA, transcribed from the *E. coli* plasmid ColE1, invades the DNA template within the origin element to form an R-loop[184]. Subsequently, RNase H cleaves the RNA moiety which is then used as a primer by DNA polymerase I[185]. In the absence of RNase H, the displaced non-transcribed ssDNA strand serves as a template for lagging-strand synthesis[184,186]. Moreover, the displaced ssDNA moiety allows a mobile primase/helicase to promote replication fork progression by providing primers for Okazaki fragment synthesis[186]. Importantly, DNA synthesis fails to start in the absence of RNA hybridization to the DNA template. These findings suggest that R-loop formation is crucial for the initiation of DNA replication. A similar mechanism for R-loop-dependent start of DNA replication has been found during early bacteriophage T4 infection for the *ori(uvsY)* and the *ori(34)* origins, during which two genomic elements participate in replication initiation[187]. First, a T4 middle-mode promoter triggers transcription of an RNA that afterward forms an R-loop in a downstream DNA unwinding element[187]. Stable R-loop formation in return promotes replication initiation by providing a primer for leading-strand synthesis in a similar fashion as in *E. coli*[187,188]. Activation of mitochondrial DNA replication is also based on the formation of an R-loop and is conserved from *Saccharomyces cerevisiae* to humans[188]. Mitochondrial DNA contains two distinct origins of replication. One for heavy strand synthesis, which is first initiated, and the other for light strand synthesis[189]. Two distinct features are required for the efficient start of mitochondrial DNA replication. One is the major promoter for mitochondrial RNA transcription, which is situated directly upstream of the site of replication initiation, the second is a GC-rich element downstream of the promoter. Transcription from this

promoter leads to the formation of an mtRNA polymerase-dependent R-loop within the GC-rich cluster[190]. *In vitro* transcription experiments revealed that this short RNA-DNA hybrid is a sufficient *cis*-element to initiate mitochondrial replication[188,190].

Not only prokaryotic and mitochondrial DNA replication can be initiated in an R-loop-dependent-manner. An up to 85 % overlap between DNA replication initiation sites and transcriptional units has been observed in mammalian cells[188,191]. Genome-wide sequencing approaches revealed that highly efficient and conserved replication origins in mouse and human cells overlap with CGIs, which are prone to R-loop formation[188,191]. This raises the possibility of an R-loop-driven mechanism for the initiation of DNA replication in mammalian cells. Intriguingly, the human ORC complex has been shown to preferentially bind G4-forming RNA and ssDNA structures over dsDNA *in vitro*[192,193]. NMR-based approaches have revealed that the human ORC1 binds to the external G-tetrad planes of the G4 structures[193]. Since almost 60% of metazoan replication origins contain G4 signatures, R-loop-stabilized G4s could attract the ORC complex to replication initiation sites to assemble the pre-replication complex[194,195]. In *Drosophila*, in which most CGIs are un-methylated, replication origins are also highly associated with CGIs, indicating that DNA replication initiation at CGIs is independent of their methylation status[194]. Recently, additional evidence for origin G-rich repeated elements (OGRE) being crucial for the initiation of DNA replication has been provided. Alterations of G4 signatures lead to the loss of origin activity at a particular locus, whereas insertion of an OGRE leads to the generation of a novel replication origin[196,197]. In conclusion, these findings suggest that replication origins can be activated by G4-forming sequence signatures within CGIs in metazoan systems. Further studies are required to decipher whether the G4s are dependent on R-loop formation at CGIs.

1.7 R-loop-dependent genomic instability

R-loops do not only contribute to cell survival as regulatory elements but can also pose a threat to genomic stability. The most prominent events that lead to R-loop-dependent DNA damage are conflicts between the replication and the transcription machinery. Since DNA synthesis and RNA transcription take place on the same template, transcription-dependent R-loop stabilization can constitute an obstacle for the DNA replication machinery. Consequent transcription-replication conflicts (TRCs) can induce replication fork stalling. If conflicts are not faithfully resolved and replication forks are not properly stabilized, as is the case in *BRCA1/2*-deficient cells, persistent stalling can lead to dissociation of the replication fork[198]. The collapse of replication forks results

in unfinished intermediates that frequently result in DSB formation, the most severe form of DNA damage. Unprotected, persistently stalled replication forks are also prone to digestion through nucleases such as MRE11[32,199,200]. Moreover, MUS81 and RECQ5 process damaged replication forks, leading to endogenous DNA damage[200].

Accordingly, global accumulation of R-loops leads to the activation of the main DDR kinases ATM and ATR and activation of the cell cycle checkpoint[172,201]. Recently, it has been shown that the directionality of the TRCs plays an important role in the severity of the conflict and the consequent genomic instability. TRCs can occur in two different orientations. The transcription bubble and the replication fork travel either co-directionally (CD) or encounter each other in a head-on conflict (HO)[172]. In bacteria, HO conflicts are more deleterious than CD conflicts by impeding replication fork progression, leading to genomic deletions and recombination events[172,202–204]. To prevent HO conflicts within the circular genomes, replication and transcription are organized co-directionally under basal conditions[205]. Strikingly in human cells, R-loops but not unperturbed transcription complexes lead to orientation-specific DSB formation and consequent DNA damage response[172]. Replication forks that encounter R-loops in a HO conflict lead to R-loop stabilization, whereas CO conflicts result in reduced R-loop levels (Figure 3). During HO conflicts, the resulting replication stress is further potentiated by de-regulated origin firing and consequent increase in HO conflicts and increased DSB formation[172].

In addition to TRCs, R-loops can alter genomic integrity due to exposure of the displaced ssDNA moiety, which is more susceptible to mutagenic recombination events compared to dsDNA[206]. The displaced ssDNA within the R-loop provides an excellent substrate for the actions of DNA modifying enzymes such as the activation-induced cytidine deaminase (AID)[122]. Once ssDNA stretches are exposed through R-loop formation, AID can associate with the target ssDNA and trigger the deamination of cytosines to uracils (Figure 3)[122,123]. The dC to dU conversion is highly mutagenic since during the subsequent S-phase the resulting U:G mismatch can lead to a C-to-T transition mutation in one of the daughter strands[88]. Once the U:G mismatch is recognized before or during the following round of DNA replication, the base-excision repair (BER) machinery will be recruited. BER components like the abasic endonuclease (APE) can process the mis-incorporated uracil, leading to a nick or an abasic site[88]. DNA nicks are prone to be converted to DSBs and ultimately trigger genomic instability. The abasic site is mutagenic since replication over the lacking nucleotide will lead to the incorporation of a random nucleotide during the following

replication cycle[88]. Therefore, the ssDNA stretch within R-loops is prone to DNA damage through DNA modifying enzymes such as AID. However, AID is only highly expressed in activated B cells, therefore this mode of R-loop-dependent genomic instability is suggested to be rare in other tissue. In cancer cells, gene expression changes might lead to upregulation of AID and therefore open the possibility of R-loop-dependent DNA damage through AID activity.

Furthermore, deleterious processing of R-loops by DNA repair proteins can result in genomic instability. Persistent R-loops that are stabilized in the absence of RNA processing factors or after inhibition of TOP1 are processed by the transcription-coupled nucleotide excision repair (TC-NER) machinery, thereby promoting R-loop-dependent genomic instability[207]. Tethering of RNA Pol II to chromatin after stable R-loop formation triggers CSB-dependent recruitment of the TC-NER factors XPG and XPF. These endonucleases recognize the R-loop and introduce a single-strand break, which can subsequently be transformed into a DSB. Alternatively, two incisions on opposing strands of the R-loop are generated, leading to direct DSB formation[207]. Intriguingly, CSB but not XPC is required for R-loop processing, suggesting that TC-NER but not global NER promotes R-loop-dependent DSB formation. A similar mode for TRC-independent induction of DNA damage has been described in non-replicating cells. Transcription blockage by topoisomerase I cleavage complexes (TOP1ccs) in quiescent human cells has been shown to recruit the endonucleases XPF and XPG or FEN1[208]. Dual incision in opposing strands of the R-loop results in DSB formation independent of the replisome or actions of DNA editing enzymes (Figure 3)[208].

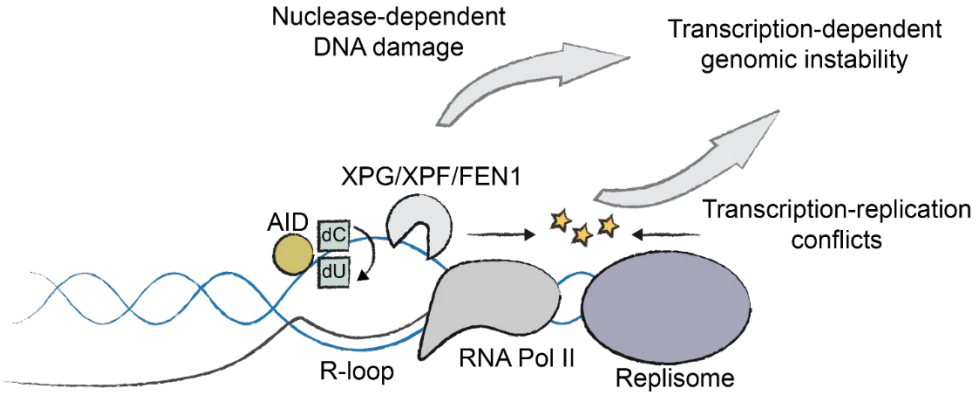


Figure 3: Excessive R-loop formation induces genomic instability through distinct mechanisms

Schematics displaying three potential mechanisms for R-loop-dependent genomic instability. DNA is depicted in blue, RNA in black, the replisome in purple, RNA polymerase in grey, activation-induced cytidine deaminase (AID) in yellow, XPG/XPF/ FEN1 nucleases in light grey. Stars indicate a transcription-replication conflict.

1.8 R-loop homeostasis is maintained by a dedicated protein machinery

1.8.1 Deleterious RNA processing results in R-loop stabilization

Most RNA processing in human cells is initiated co-transcriptionally, while the transcript is still conjugated to the RNA polymerase. Therefore, alterations during nascent RNA processing have been linked to excessive R-loop formation by tethering the transcription complex to the DNA[209]. A major component of eukaryotic RNA maturation is splicing, during which intronic sequences within the newly synthesized transcript are coordinately excised, resulting in an exon-dominated RNA. Loss of function mutations in splicing factors and causal splicing alterations lead to increased RNA polymerase pausing[209,210]. Due to reduced elongation, the transcription complex can be tethered to chromatin by R-loop stabilization. Splicing factor mutations are frequently identified in patients suffering from myelodysplastic syndromes (MDS)[211–213]. Intriguingly, RNA-sequencing analysis revealed that barely any commonly altered splicing events are identified between mutations of different splicing factors while resulting in the same pathogenesis[210]. These findings suggest an alternative unifying mechanism, independent of individual splicing events. Recently, it has been proposed that expression of MDS splicing factor variants of SRSF2 and U2AF1 results in augmented R-loop formation due to increased RNA Pol II pausing[210]. As a result, replication-dependent DNA damage triggers checkpoint activation, which leads to reduced proliferation of hematopoietic progenitors[210]. How splicing impairments mechanistically result in R-loop formation remains elusive. However, a recent study suggests that the RNA helicase DHX9 promotes R-loop formation in the absence of core splicing factors[209]. DHX9 associates with RNA Pol II and co-transcriptionally unwinds secondary structures within the nascent transcript. Loss of SFPQ or SF3B3 results in R-loop-dependent tethering of RNA Pol II to chromatin, which is dependent on the presence of DHX9. Impaired removal of secondary structures within the nascent transcript in the absence of DHX9 repels the spliceosome and thereby prevents splicing initiation. In contrast, inhibition of the SF3B splicing complex by Pladienolide B (PladB) and consequent intron retention has been reported to result in a genome-wide loss of R-loop formation through a global reduction in nascent transcription[69]. Only a subset of genes displayed R-loop gains due to loss of transcriptional termination and read-through transcription. These findings highlight that the simple hypothesis according to which splicing alterations result in R-loop stabilization is not valid. Even though transcription dynamics, splicing, and R-loop formation correlate, further studies are required to elucidate the co-dependencies of the different processes.

1.8.2 The role of DEAD/H-box helicases in R-loop regulation

Various studies implicate DEAD/H-box helicases as R-loop resolvases in human cells. However, the links between depletion of a DEAD/H-box helicase and changes in R-loop levels remain mechanistically poorly characterized. One of the DEAD-box helicases is DDX18, which has been shown to associate with RNA-DNA hybrids *in vitro*[214]. Incubation with nuclear lysates led to significantly higher binding of DDX18 to RNA-DNA hybrid oligonucleotides compared to a dsDNA substrate. The *in vitro* binding affinity was validated by biolayer interferometry using dsDNA and RNA-DNA substrates together with purified DDX18 protein[214]. Co-localization of DDX18 with nucleolin and the S9.6 antibody served as *in vivo* proof that DDX18 acts as an R-loop regulator. Another helicase that has been shown to regulate R-loop levels is DDX21 whose helicase activity towards RNA-DNA hybrids is dependent on its acetylation status[215]. De-acetylation of DDX21 by SIRT7 significantly promotes the *in vitro* unwinding activity and loss of either DDX21 or SIRT7 results in a global R-loop upregulation in MCF7 cells. Post-translational modification (PTM)-dependent regulation has also been observed for the DDX5 helicase. PRTM5 methylates DDX5 at its RGG/RG motif and thereby allows the association of XRN2 to DDX5[216]. Together, XRN2 and DDX5 control R-loop homeostasis at transcription termination sites downstream of the poly(A) signal, thereby promoting faithful release of RNA Pol II. Moreover, the nucleopore-associated helicase DDX19 has been reported to resolve R-loops in response to replication stress and DNA damage. DDX19 re-localizes from the nuclear periphery to the nucleus after various types of DNA damage (UV, APH, CPT) in an ATR-CHK1-dependent manner[217]. In the absence of DDX19, replication-transcription conflicts are not resolved and cells suffer from R-loop-dependent DNA damage. The K-homology domain-containing helicase DDX43 has been reported to unwind RNA-DNA hybrids *in vitro*[218]. However, also dsRNA was unwound by purified DDX43, suggesting unspecific unwinding activity. Furthermore, no data regarding *in vivo* R-loop regulation of DDX43 has been published so far. Recently, a role as an R-loop resolvase for the UAP56/DDX39B helicase has been reported. DDX39B associates with actively transcribed genes and globally opposes deleterious R-loop accumulation across the genome[219]. The role of the DHX9 helicase in R-loop regulation is still quite controversial. DHX9 has been reported to regulate R-loop levels at transcription termination sites to support the XRN2-dependent release of RNA Pol II[152]. In addition, DHX9 has also been implicated in regulating R-loops in the promoter region in a TDRD3/TOP3B-dependent manner[220]. Intriguingly, S9.6 dot blot analysis within the same publication revealed that global R-loop levels are reduced after knockdown of DHX9, suggesting

that DHX9 does not oppose R-loop formation. This observation is in line with Chakraborty´s report that suggests an R-loop promoting function of DHX9 rather than a preventing function[209]. Furthermore, DHX9 is linked to a role in DNA repair by promoting RNA-dependent BRCA1 recruitment to DSBs during HR[221]. The association in DSB repair makes the interpretation of R-loop level changes in the absence of DHX9 even more complex, since DSBs also promote R-loop formation[168].

1.8.3 Ribonucleases H1 and H2 – efficient regulators of R-loops

Two main players in removing RNA-DNA hybrids from chromatin are the ribonucleases RNaseH1 and H2[222,223]. RNase H enzymes were firstly identified in 1969[224]. RNaseH1, which consists of three distinct domains, is a highly structured endonuclease that can specifically hydrolyze RNA within RNA-DNA hybrids[225,226]. A mitochondrial localization signal (1-26) precedes the hybrid binding domain (HBD) (27-76) within the N-terminus. The catalytic domain at the C-terminus is conjugated to the HBD via a linker domain. RNaseH1 possesses a strong affinity towards RNA-DNA hybrids via its HBD, which highly discriminates between hybrids and other nucleic acid substrates[226]. Overall, RNaseH1 is positively charged, except for the substrate interaction surface which is neutral to acidic, thereby contributing to substrate specificity[225]. The HBD within human RnaseH1 possesses a DNA binding channel formed by a basic protrusion alongside a phosphate-binding pocket[225]. RNA specificity is provided by four adjunct 2´-OH groups, thus preventing catalytic activity towards substrates smaller than 4 ribonucleotides[223]. Cleavage of the RNA moiety from the RNA-DNA hybrid is dependent on two divalent metal ions, such as Mg^{2+} [225]. Despite the protein structure, very little is known about the regulation of RNaseH1 activity and substrate specificity in human cells. Association of RPA with the ssDNA moiety within R-loops and interaction with RNaseH1 have been shown to enhance the catalytic activity of RNaseH1 *in vitro*[227]. RPA binding-defective mutants abrogate the R-loop suppressing function of RNaseH1 *in vivo*[227]. Various roles of RNaseH1 in eukaryotic R-loop regulation have been described. RnaseH1 null mice embryos arrest during development at E8.5[228]. The growth defect is caused by a lack of mitochondrial RNaseH1, which is required for mtDNA replication by removing RNA primers. The subsequent loss of mitochondria in RnaseH1 null mice embryos results in embryonic lethality[228]. Similar to loss of function of RnaseH1, increased activity of RNaseH1 also leads to aberrant mitochondria. Expression of a pathological variant that is linked to mitochondrial disease leads to increased R-loop removal in mitochondria and subsequently to mtDNA aggregation since RNA-DNA hybrids promote physical segregation of compact mtDNA[229]. In addition to its mitochondrial

function, RnaseH1 is also required to maintain genomic stability in the nucleus. Human RNaseH1 ensures faithful DNA replication fork progression by degrading excessive R-loops, thereby preventing collisions with stalled transcription machinery[230,231]. Moreover, the removal of RNA-DNA hybrids at DSBs is crucial for faithful DNA repair[168]. However, the R-loop regulatory function of RNaseH1 is dispensable for efficient DSB repair, indicating that it does not play a role in removing DSB-associated RNA-DNA hybrids[231].

Other than RnaseH1, RNaseH2 has evolved quite rapidly from a monomeric enzyme in prokaryotes to a heterotrimeric complex in eukaryotes[223]. The human RnaseH2 active complex consists of three subunits RNaseH2A, B and C. Pathogenic germline mutations in either of the three subunits is predispose patients to Aicardi-Goutières syndrome (AGS), an inflammatory-dependent neurodevelopmental disorder[232]. In contrast to RnaseH1, which can only act on substrates above three ribonucleotides, the H2 complex can bind and cleave a single ribonucleotide within a DNA duplex[233]. Thus, RNaseH2 participates in the removal of mis-incorporated ribonucleotides by DNA polymerase during DNA replication[234,235]. Interaction of PCNA with RnaseH2B subunit via its PIP domain enables RNaseH2 to associate with replication forks, at which RNaseH2 removes Okazaki fragments during lagging strand synthesis[223]. Removal of ribonucleotides by RnaseH2 generates DNA containing 3′-hydroxyl and 5′-phospho-ribonucleotide ends, which constitutes an optimal target for the Okazaki fragment maturation machinery[16,223,234–236]. Very little is known about the regulation of the RnaseH2 complex in human cells. In yeast, RnaseH2 activity is restricted to certain cell cycle phases in contrast to RnaseH1, which acts independently of the cell cycle stage[237]. However, further studies are required to elucidate if similar cell cycle-dependent regulation is evident in human cells.

1.8.4 m^{6}A-modified RNA within R-loops influences their stability

Little is known about specific DNA or RNA modifications within R-loops. In eukaryotes, one of the most frequent modifications within RNA is N^6-methyladenosine (m^{6}A), which affects various steps of post-transcriptional RNA metabolism[238]. Recently, a subset of R-loops containing m^{6}A-modified RNA has been identified, which suggests that some R-loops are specifically m^{6}A-modified while others are not[239]. Furthermore, m^{6}A deposition within R-loops is cell cycle-dependent. The highest levels of m^{6}A positive R-loops appear during the G_2/M phase, while m^{6}A containing R-loops are lost during G_0/G_1[240]. Currently, there is a strong discrepancy in the field of how m^{6}A-modified RNA impacts R-loop stability. On the one hand, loss of the main m6A writer,

METTL3, has been described to lead to reduced R-loop levels at the transcription termination site, suggesting that m^{6}A modification promotes R-loop formation[239]. On the other hand, m^{6}A deposition within R-loops has been reported to lead to degradation of the structure[240]. R-loop destabilization is dependent on the m^{6}A reader protein YTHDF2, which associates with R-loop-rich regions within the genome. Loss of YTHDF2 leads to R-loop accumulation and spontaneous DNA damage, suggesting a role for m^{6}A in safeguarding the genome against R-loop-dependent genomic instability. In line with the idea that m^{6}A modifications serve as R-loop resolution signals, the tonicity-responsive enhancer binding protein (TonEBP) promotes m^{6}A-dependent R-loop destabilization[241]. TonEBP recognizes UV and CPT-induced R-loops and preferentially binds them *in vitro* compared to DNA. *In vivo*, TonEBP binds METTL3 and promotes m^{6}A deposition at R-loops and subsequent recruitment of RNaseH1. Together, the two studies suggest that m^{6}A serves as a signal to trigger downstream R-loop resolution pathways.

1.9 Proteomic studies to identify R-loop interacting proteins

As of now, two approaches have been published that map R-loop-associated proteins. Wang and colleagues used *in vitro* RNA-DNA hybrid oligonucleotides to identify R-loop binding proteins from human nuclear extracts[214]. Roughly 800 proteins were associated with RNA-DNA hybrids of which 300 proteins displayed a stronger affinity towards hybrids compared to a dsDNA control. Gene set enrichment analysis revealed that proteins involved in RNA binding, mRNA splicing, and ATP-dependent helicase activity associate with RNA-DNA hybrids. Alpha-beta plait, P-loop triphosphate hydrolase, DEAD/H-box, nucleic acid-binding, OB-fold, and K-homology are domains that are significantly enriched among RNA-DNA hybrid binding proteins. Furthermore, the affinity of DDX18 to RNA-DNA hybrids was validated *in vitro*.

The second proteomic screen facilitated the S9.6 antibody to co-immunoprecipitate RNA-DNA hybrids and associated proteins from nuclear extracts after chromatin extraction and fragmentation[152]. RNA binding proteins, DNA binding proteins, mRNA processing factors, and histones were among the proteins interacting with affinity-purified R-loops. Of note, the identifications were not sensitive to RnaseH1 treatment. However, enrichment of *in vivo* nuclear R-loops and re-probing with nuclear extracts led to similar identifications as the initial Co-IP. In particular, the DEAH-box helicase DHX9 was identified as a strong R-loop interactor and its role in R-loop-dependent transcriptional termination was functionally characterized.

1.10 Genome-wide R-loop mapping approaches

The classical approach to map R-loops genome-wide is called DNA-RNA hybrid immunoprecipitation (DRIP). It is based on affinity purification of RNA-DNA hybrids using the S9.6 antibody. After cell lysis and purification of nucleic acids, a cocktail of restriction enzymes is used to digest the DNA into smaller fragments to allow successful enrichment. Half of the input is usually digested with RNAseH1 *in vitro* to validate the specificity of the S9.6 antibody towards RNA-DNA hybrids. Affinity purified nucleic acids are subsequently analyzed by Next-generations sequencing (NGS) or quantitative polymerase chain reaction (qPCR). Recently, small changes in the workflow have been introduced to allow strand-specific R-loop mapping[151]. The strand-specific protocol termed DRIP-c is similar to classic DRIP using roughly 5 times the input material. After S9.6-based affinity purification, DNA is depleted using a DNAse. The remaining RNA is reverse-transcribed to cDNA and subsequently analyzed by NGS. Since only the RNA moiety within the hybrid and not the DNA component are sequenced, the signal coming from DNA surrounding the hybrid is omitted. Thereby, strand-specificity is gained and a higher resolution in respect of the position of the hybrid is achieved.

Since the S9.6 antibody not only binds to RNA-DNA hybrids - but to a smaller extent - also to RNA species, alternative antibody-independent approaches have been developed. R-ChIP was the first workflow that allowed RNA-DNA hybrid affinity purification and subsequent sequencing without the S9.6 antibody[96]. Catalytically dead RNaseH1 (D210N) carrying a V5 tag is exogenously expressed in cells. D210N-RNaseH1 retains the ability to bind RNA-DNA hybrids while losing its catalytic activity and thus getting trapped on the substrate. Affinity purification with a V5 antibody after crosslinking D210N-RNaseH1 to chromatin allows enrichment of R-loop forming regions and subsequent NGS analysis. The advantage of R-ChIP is the independency of the S9.6 antibody. Limitations are the restriction to the identification of R-loops that are bound by RNaseH1 and the lack of RnaseH1 digestion control. In addition, cells need to be genetically modified to express the bait which prevents R-loop mapping in primary material.

Another S9.6-independent approach is called MapR. The core of the CUT&RUN-based workflow is a purified catalytically-dead RNase H, which is conjugated to an MNase[156,242]. Under native conditions, the nuclease is targeted to R-loops via the catalytically-dead RNase H. Activation of the MNase allows local fragmentation of chromatin occupied by R-loops. Subsequently released fragments are purified and analyzed by NGS. The advantages are a high signal-to-noise ratio,

identification of R-loop containing regions under native conditions, and very low input. The limitation is, as for R-ChIP, that MapR can only identify R-loops that are bound by RNase H.

1.11 R-loop dysregulation contributes to neurological disorders and cancer

Dysregulation of R-loops is a driver of genomic instability, one of the hallmarks of cancer[243]. Mutations in R-loop-regulating genes are highly associated with the development of different cancers. Germline and somatic *BRCA1* mutations are frequently identified in breast and ovarian cancers[244–246]. Besides its prominent role in HR and replication fork protection, BRCA1 has recently been described as an R-loop regulator at RNA Pol II transcription termination sites and telomeres[126,247]. BRCA1 suppresses deleterious R-loop formation and thereby prevents spontaneous DNA damage. Sequencing of 21 breast cancers revealed that patients with somatic *BRCA1* mutations carried more insertion/deletions in the proximity of R-loop positive terminator regions than patients with *BRCA2* mutations or without *BRCA1/2* mutations[247,248]. These correlations suggest that R-loop dysregulation in the presence of pathogenic *BRCA1* mutations contributes to cancer development.

The *SETX* gene encodes the senataxin RNA/DNA helicase, which is a major R-loop regulator in human cells[101,207]. Germline mutations in *SETX* predispose patients to neurological disorders such as ataxia with oculomotor apraxia type 2 (AOA2) and juvenile amyotrophic lateral sclerosis (ALS4)[249,250]. R-loop homeostasis is disturbed in cells of patients with pathogenic *SETX* germline mutations. Since SETX acts as an R-loop resolvase, it has been hypothesized that R-loops would be more stable in ALS4 patients. However, R-loops are globally downregulated in cells from patients with pathogenic *SETX* mutations[163]. Reduced R-loop occupancy of the *BAMBI* promoter leads to increased binding of DNMT1, which subsequently methylates the *BAMBI* promoter. Silencing of BAMBI leads to activation of the TGF-ß pathway, which alters neuronal differentiation in an ID1/2-dependent manner. Genome-wide analysis revealed methylation alterations of 1,200 gene promoters in ALS4 cells with pathogenic *SETX* mutations. These findings suggest a contribution of R-loop dysregulation to ALS4 development.

Somatic mutations in splicing factors such as SF3B1, U2AF1, and SRSF2 are frequently identified in patients with myelodysplastic syndromes (MDSs)[212]. In addition to alterations during hematopoiesis and dysplastic differentiation of hematopoietic stem and progenitor cells (HSPCs), MDS patients possess a strong predisposition to acute myeloid leukemia (AML)[251]. The high

frequency of splicing factor mutations in MDS patients suggests that a commonly altered splicing event drives the disease[210,252,253]. However, next-generation sequencing of MDS patients revealed unique splicing alteration patterns for each mutation. Since a common feature of MDS and AML is increased genomic instability[254], Chen et al. investigated the cause of DNA damage in cellular models[210]. Expression of SRSF2 and U2AF1 MDS mutants leads to elevated R-loop levels due to impaired pause release of RNA Pol II[210]. Moreover, R-loop accumulation resulted in replication stress and activation of the ATR-CHK1 pathway. Rescue of the reduced proliferation of bone-marrow-derived blood progenitors by overexpression of RNaseH1 further demonstrated that augmented R-loops contribute to the MDS phenotype.

1.12 R-loops can be distinguished into two distinct classes

There are two main groups of genome-wide R-loop sequencing approaches. One group facilitates the RNA-DNA hybrid S9.6 antibody, while the other group uses catalytically-dead RnaseH1 as a probe. RnaseH1-based methods, such as R-ChIP map approximately 60% of R-loops to promoter-proximal regions, 17% to gene bodies, and 6% to transcription termination regions[255]. In contrast, S9.6-based approaches such as DRIP map 13% to promoter-proximal regions, roughly 50% to gene bodies, and 19% to transcription termination sites[89]. Of note, the R-ChIP signal declines abruptly behind the transcription start site (TSS), while R-loops identified by DRIP reach up to 1.5 kb behind the TSS[72]. Moreover, RNaseH1-based sequencings revealed R-loop formation in intragenic enhancer regions and tRNA genes[255], whereas very little signal from these genomic loci is obtained by DRIP[72]. These findings suggest that the two groups of R-loop mapping approaches identify alternative classes of R-loops. However, *in vivo* R-loop mapping using the S9.6 antibody led to similar results as previous RnaseH1-based mappings[256]. *Ex vivo* R-loop mappings using RnaseH1-based methods identify similar R-loop landscapes as previous S9.6-based approaches[256]. In accordance, not the probe for detection but rather the experimental setup defines the class of R-loops that is identified. Recently, a first attempt to distinguish the two R-loop classes based on their properties has been made. Promoter-proximal R-loops, which are detected by *in vivo* mapping strategies are termed Class I R-loops, whereas gene body R-loops, which are detected *ex vivo* are defined as Class II R-loops[72]. Class I R-loops are short (up to 60 bp), while Class II R-loops are long (median length of 1500 bp)[72]. Interestingly, Class II R-loops consist of individual R-loops with a median length of 300 bp, which cluster together within R-loop hotspots and appear as one peak during *ex vivo* mappings[257]. The high correlation between RNA pol II ChIP-sequencing, nascent RNA sequencing, and Class I R-loops suggests that R-loops belonging to this class are

strongly associated with paused RNA Pol II[258,259]. Mechanistically, both R-loop classes might contribute to slowing or pausing of RNA Pol II: (1) Class I R-loops contribute to promoter-proximal pausing; (2) Class II R-loops support intragenic slowdown or pausing at the transcription termination site[72].

1.13 The tumor suppressor DEAD-box helicase 41

DEAD-box helicase 41 (DDX41) is a member of the DEAD/H family and is essential in human cells. DDX41 is a tumor suppressor that is conserved in *D.melanogaster*, *C.elegans*, *D.rerio,* and plants[260–263]. Somatic and germline mutations in *DDX41* are drivers for adult myelodysplastic syndrome (MDS)/acute myeloid leukemia (AML). Pathogenic variants of DDX41 are frequently evident in 0.5% to 4% of MDS/AML patients[264]. Germline DDX41 mutations mostly lead to frameshifts that result in the expression of a truncated protein[265]. However, somatic mutations predominantly occur within the DEAD-box or helicase domain, suggesting to cause altered helicase activity[265]. Taken together, disease development through pathogenic *DDX41* mutations most likely results from loss of function or reduced function of DDX41. Structural analysis of DDX41 revealed that its structure can be divided into four elements (Figure 4)[266]. The N-terminus is unstructured and harbors a putative nuclear localization signal. ATP binding and nucleic acid binding are facilitated by the conserved DEAD-box domain, which consists of seven motifs (Q, I, IA, IB, IC, II (DEAD), and III). Adjunct to the DEAD-box is the RecA-like helicase domain, consisting of four motifs (IV, IVA, V, and VI), which is followed by a Zinc finger in the C-terminal domain[266–268]. Deletion of DDX41 leads to mRNA splicing defects and compromised RNA processing[269,270]. Moreover, DDX41 interacts with components of the spliceosome, thus a role for DDX41 in mRNA splicing has been proposed[270]. As applies to the human DDX41, the *C.elegans* orthologue SACY-1 also interacts with proteins involved in RNA splicing[262]. Loss of SACY-1 leads to RNA processing defects through splicing-dependent and splicing-independent mechanisms. The zebrafish variant Ddx41 has recently been proposed to suppress excessive R-loop formation, thereby preventing cGAS-STING-dependent inflammatory response. Activation of the inflammatory pathway in the presence of Ddx41 mutants results in the expansion of HSPCs, providing a link to the development of MDS/AML in patients with pathogenic DDX41 variants. Despite first evidence that loss of DDX41 leads to splicing alterations and transcriptome changes, the molecular functions of DDX41 which contribute to the development of MDS/AML remain poorly understood.

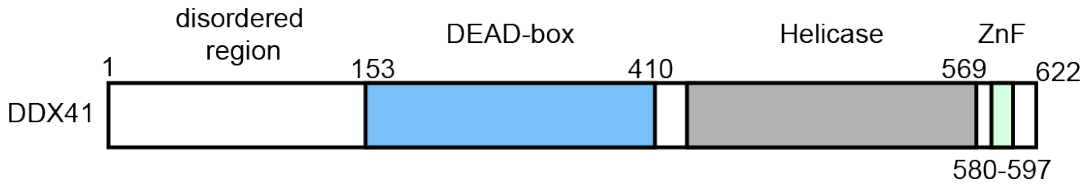


Figure 4: Schematic representation of human DDX41

Schematics displaying DDX41 domain organization. Amino acid positions are indicated by the numbers. N-terminus is shown on the left and C-terminus on the right. The blue box depicts the DEAD-box domain, grey the helicase domain, and green the zinc finger (ZnF) domain.

1.14 Excessive R-loop formation triggers DNA damage response signaling

One of the main players in the DNA damage response is the ATM kinase, which promotes the repair of DSBs[271,272]. The most prominent target of ATM is the cell cycle checkpoint kinase CHK2[273,274]. R-loop-dependent activation of ATM occurs in response to co-directional TRCs, which is apparent by increased phosphorylation of the main downstream targets Chk2, KAP1, and pRPA S4/8[172]. Moreover, R-loop accumulation in the absence of WRN helicase results in XPG-dependent ATM signaling[275]. Suppression of R-loop accumulation or depletion of XPG in Werner syndrome cells (WSCs) prevents ATM activation. In accordance, chemical inhibition of ATM leads to increased R-loop-dependent genomic instability in WSCs[275]. R-loop-dependent ATM activation has also been described in the absence of DNA replication machinery in non-replicating cells, suggesting a TRC-independent function of ATM[276]. Pausing of RNA polymerase II at transcription-blocking lesions leads to R-loop stabilization and subsequent activation of ATM that in return phosphorylates the core spliceosome to displace it[276]. Taken together, there is strong evidence that ATM is activated in response to excessive R-loop accumulation in replicating and non-replicating cells[98]. The exact molecular mechanism of how ATM is activated remains poorly understood.

ATR is another master kinase in response to DNA damage, in particular in response to replication stress[277–279]. Since R-loops constitute obstacles for the replication machinery and can cause TRCs, it has been hypothesized that ATR gets activated upon loss of R-loop homeostasis. In contrast to ATM, ATR has been reported to be specifically activated in response to HO TRCs[172]. During S-phase, the ATR-Chk1 axis is activated by deleterious R-loop formation and prevents R-loop-dependent DSB formation[201]. Activation of ATR in response to excessive R-loop levels requires fork reversal and is dependent on the MUS81 nuclease[201]. Moreover, loss of ATR causes R-loop accumulation and reduced replication fork speed, suggesting that ATR is required to maintain R-

loop homeostasis and ensure replication fork progression[280]. How ATR opposes R-loop accumulation remains unclear. However, first evidence suggests that it might recruit RNA/DNA helicases such as DDX19 to oppose excessive R-loop formation[217]. Further data needs to be provided to mechanistically understand the roles of ATM and ATR and if also other kinases are involved, in opposing R-loop-dependent DNA damage.

1.15 DNA damage can activate inflammatory signaling pathways

The innate immune response is a crucial pathway for cells and organisms to protect themselves against the invasion of pathogens and tissue damage. Inflammatory signaling needs to be tightly balanced since reduced activity can lead to infections and tumorigenesis, whereas hyperactivity and self-directed activity result in allergies and autoimmune diseases[281]. During infection, exogenous components from pathogens are recognized in cytoplasmic or intercellular regions and activate a signaling cascade that results in active inflammation[282]. In addition, the inflammatory response can be triggered by endogenous sources such as cytoplasmic nucleic acid species, which can arise from replication stress-dependent DNA damage[283–287]. Accumulation of DNA-damage-dependent ssDNA or dsDNA in the cytoplasm activates the cyclic GMP-AMP synthase (cGAS)-stimulator of interferon genes (STING) pathway[286,288]. cGAS acts as a cytosolic DNA sensor that generates cyclic guanosine monophosphate–adenosine monophosphate (cGAMP) upon binding[289,290]. Production of cGAMP leads to the oligomerization of STING, which subsequently activates the TBK1 kinase[291–293]. In return, TBK1 phosphorylates STING, which induces TBK1-dependent phosphorylation of the transcription factor Interferon Regulatory Factor 3 (IRF3)[294–296]. Phosphorylated IRF3 shuttles to the nucleus to promote the transcription of type I interferons (IFNs)[294,296–298]. Moreover, TBK1 phosphorylates p65, thereby activating the canonical NF-kB pathway[296,299,300]. Taken together, the cGAS-STING pathway triggers an inflammatory response in the presence of DNA damage-induced cytosolic nucleic acids. In addition to DNA-damage-dependent cytosolic DNA, inflammatory signaling can also be triggered by the formation of micronuclei, which are generated by aberrant mitosis due to under-replicated DNA upon replication stress[301]. Leakage of nucleic acids into the cytoplasm due to the rupture of the micronuclei envelope activates the cGAS-STING pathway as described above[302].

Persistent inflammation can have dramatic consequences for distinct cell types. Especially hematopoietic stem cells and their microenvironment are sensitive to chronic inflammation since they have to endure various types of DNA damage including reactive oxygen species, telomere

shortening, and replication stress[303]. DNA damage-induced inflammatory signaling and subsequent inflammation in the microenvironment can display adverse events in hematopoietic stem cells, which can negatively impact their proliferation, self-renewal capacity, mobilization, and differentiation[304–306]. With age, the chance of DNA-damage-dependent release of pro-inflammatory cytokines into the stem cell microenvironment is significantly higher, thereby contributing to the dysfunction of aging tissue[303,307–310]. Moreover, inflammatory signaling has been proposed to induce stem cell expansion and tumor progression[310,311].

1.16 Excessive R-loop formation triggers an inflammatory response

Global R-loop accumulation leads to replication stress and genomic instability. DNA damage and genomic instability have been described as potent activators of inflammatory pathways, especially through cGAS-STING[312]. Since dysregulation of R-loops is a major source of endogenous DNA damage, it has been hypothesized that excessive R-loop formation might trigger an inflammatory response[74]. In accordance, DNA damage-induced R-loops in aged pancreata contribute to the accumulation of ssDNA in the cytoplasm, thereby inducing a virus-like inflammatory response[313]. Importantly, delivery of RNase H into inflamed pancreatic cells and *Ercc1*$^{-/-}$ mice, via extracellular vesicles, reduces the amount of R-loops, cytoplasmic ssDNA, and chronic inflammation[313]. Moreover, excessive R-loop formation in zebrafish has been shown to lead to a cGAS-STING-dependent inflammatory response that alters the numbers of hemogenic endothelium and HSPCs[260]. Maintaining R-loop homeostasis and preventing R-loop-dependent inflammatory signaling is crucial to ensure faithful hematopoiesis. Further studies are required to elucidate the connection between R-loop dysregulation, genomic instability, cytoplasmic nucleic acids, and inflammatory signaling.

1.17 Proximity labeling-based proteomics

Classical approaches to studying protein interactions depend on the extraction of proteins from their native environment. The two main families of proximity-based labeling strategies are BioID and APEX (Figure 5). Both methods and their respective variations allow the identification of protein interactions within their native context. This is crucial for mapping protein interactions in cellular spaces where intact compartmentalization is of utmost importance, e.g. chromatin environment. Cell lysis during conventional Co-IP studies leads to the disruption of chromatin context and thereby to the identification of artificial protein interactions which are not formed under native conditions. Moreover, proximity labeling approaches allow the identification of transient

interactions and high-stringency washes during the purification due to the strong affinity between biotin and streptavidin[314]. BioID-based strategies rely on the activity of a 35 kDa *E.coli*-derived BirA* biotin ligase which catalyzes the reaction of biotin and ATP to biotinoyl-5′-AMP[315]. The reactive biotin derivative subsequently diffuses and quickly reacts with available primary amines of lysines within proximal proteins. The introduction of a mutation (R118G) within the catalytic region of the BirA* ligase allows sudden release of the biotinoyl-5′-AMP molecule after activation[315]. Several adaptations of the first-generation BirA* enzyme have been developed over time. BioID2 is based on a smaller 27 kDa biotin ligase from *Aquifex aeolicus* with an R40G mutation in the catalytic domain[316]. Although BioID2 is smaller and requires less biotin than the first-generation BioID, both enzymes display comparable activity[316]. Independently, two further variations of the *E.coli*-derived BirA* have been generated based on yeast surface display screenings[317]. TurboID (35 kDa) and miniTurbo (28 kDa) contain 14 and 12 mutations in the RBAM domain, respectively, which reduce the labeling time from hours to minutes[317]. The second family of proximity labeling techniques is based on the activity of the ascorbate peroxidase APEX which generates a biotin-phenoxyl radical in the presence of biotin-phenol and H_2O_2[318]. Subsequently, the activated biotin-phenoxyl radical reacts quickly with neighboring electron-rich amino acids such as tyrosine, tryptophan, cysteine, and histidine[319]. Thus, APEX only requires short reaction times (< 2 minutes). APEX has been engineered by introducing three mutations (K14D, W41F, E112K) in the soybean-derived ascorbate peroxidase APX, which significantly increased its catalytic activity compared to the wild type[318]. Direct evolution of APEX by a single mutation (A134P) led to the discovery of the far more active APEX2 enzyme[320]. Proximity labeling-based proteomics has been broadly applied to identify proximal protein networks in plants[321], xenograft tumors[322], embryonic stem cells[323], and compartments within human cells such as the mitochondria[324,325] and ER membranes[326].

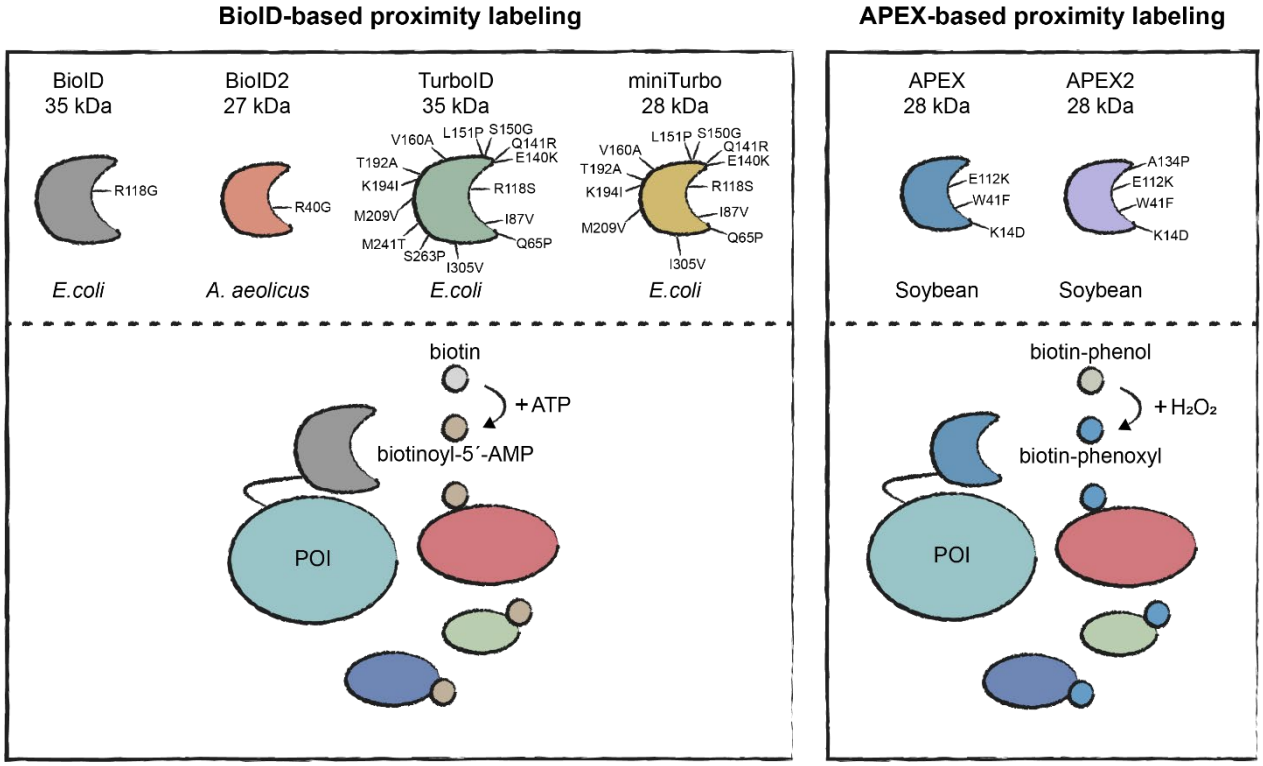


Figure 5: BioID or APEX-based proximity labeling approaches

Schematics displaying different proximity labeling approaches. BioID-based on the left and APEX-based on the right. Different variations of the used enzymes are color-coded. Mutations from the wild type are indicated. Name and molecular weight are stated above, origin below. The two enzymatic reactions are depicted on the bottom. POI = protein of interest.

1.18 Quantitative mass spectrometry (MS)-based proteomics

1.18.1 Shotgun proteomics

MS-based "bottom-up" proteomics defines the analysis of proteins after proteolysis into peptides[327]. Inspired by the term for shotgun genomic sequencing, shotgun proteomics refers to the "bottom-up" identification of individual proteins from a heterogeneous protein mixture[328,329]. After extraction from their cellular context, proteins are digested into peptides by sequence-specific proteases such as trypsin which is commonly used for shotgun proteomics. Trypsin cleaves C-terminal to arginine and lysine residues, thus generating peptides that contain a positive charge which is advantageous for downstream MS analysis[330]. Depending on the sample complexity and peptide length, a combination of selective and unselective proteases can increase the protein and PTM coverage[331]. Fractionation of proteins by SDS-PAGE before proteolytic digestion can further increase the coverage by reducing sample complexity and removing interfering chemicals. Separation of peptides by reversed-phase high-performance liquid chromatography (RP-HPLC) is commonly used to further reduce the complexity of the peptide mixture prior to MS analysis.

1.18.2 Peptide sequencing by LC-MS/MS

Even though current mass spectrometers provide high resolution and speed, complex peptide samples need to be fractionated through RP-HPLC to achieve high peptide identification rates. Commonly, peptides are separated based on their hydrophobic properties using octadodecyl alkane (C_{18})-packed columns. Different retention times of peptides on the C_{18} stationary phase of the column in a gradient with increasing organic solvent in the mobile phase are exploited to separate peptides before MS analysis[332]. Identification by MS requires ionization of the peptides after chromatographic resolution. Therefore, peptide mixtures are commonly subjected to MS analysis at low pH conditions. Electrospray ionization (ESI) is usually combined on-line with liquid chromatography for peptide MS analysis[333,334]. During ESI, the sample leaves the RP-HPLC through a thin capillary which promotes the formation of droplets containing the peptides[335]. An electric field between the tip of the capillary and a counter electrode causes the positively charged droplets to fly towards the MS inlet. While traveling towards the negative electrode, liquid evaporates from the positively charged droplets until the Coulomb force outgrows the surface tension, resulting in a Coulomb explosion[336]. Very small droplets are formed that experience complete solvent evaporation. As a result, positively charged peptide ions enter the mass spectrometer and migrate towards the mass analyzers and the detector[337,338]. Mass analyzers, such as quadrupoles and Orbitraps can determine the mass-to-charge ratio (m/z) of incoming ions. Mass spectrometers such as the Q Exactive plus (used in this study) possess a combination of mass analyzers for MS/MS analysis. The quadrupole is facilitated as a mass filter that only allows passage of a defined m/z ratio, thereby isolating ion precursors. It consists of four parallel rods which generate an oscillating electric field by switching from positive to negative charge[339]. Setting a specific voltage allows ions with a defined m/z ratio to pass the quadrupole in trajectories, while molecules with a deviant m/z ratio are discharged by colliding onto the rods[340]. Orbitraps, which can be used as mass analyzers and detectors, consist of a center electrode that is enclosed by two outer electrodes[341]. Ions that enter the Orbitrap are trapped and oscillate in the inner space between the electrodes. Mass spectra are detected by Fourier transformation of the ion-induced oscillation frequency measured by image current detection on the outer electrodes[342]. Using a series of mass analyzers and a detector allows the analysis of peptides by so-called tandem mass spectrometry (MS/MS). Analysis of an ion precursor spectrum (MS1) elucidates the mass of a peptide. Isolation of a specific peptide, its fragmentation in a higher-energy C-trap dissociation (HCD) cell, and subsequent analysis of the ion fragmentation spectrum (MS2) provide the sequence information of

the chosen peptide[343]. During data-dependent acquisition (DDA) the top N precursor ions from the MS1 spectrum are selected, fragmented and the MS2 spectra generated. MS2 spectra are matched to the *in silico*-digested proteome of interest to define which peptides were present in the subjected peptide mixture[344–346]. Moreover, state-of-the-art software can assemble proteins from all identified peptides to conclude which proteins were present in the initial sample[347].

1.18.3 Stable isotope labeling with amino acids in cell culture (SILAC)

MS/MS can not only be used to identify certain peptides from a complex mixture but can also be applied quantitatively. Different quantification methods are available, ranging from label-free quantification (LFQ)[348] over chemical labeling strategies such as tandem mass tags (TMT)[349], to metabolic labeling approaches such as stable isotope labeling by amino acids in cell culture (SILAC)[350]. LFQ is easily applicable by quantitatively comparing peptide intensities between two or more samples in distinct MS runs. While no metabolic or chemical labeling of the peptides is required, LFQ suffers from limited reproducibility, since peptides are measured in separate MS runs. Quantification based on TMT-labeling is more robust since peptides are measured within the same MS analysis. Moreover, up to 16 distinct samples can be multiplexed[351]. However, TMT labeling is very expensive and suffers from co-isolation interference[352]. Co-isolation of multiple peptides for fragmentation interferes with the reporter quantification, thereby negatively impacting quantification accuracy. This study relies on MS/MS quantification based on metabolic labeling of proteins in human cells using SILAC. Cells are grown in medium containing isotopically labeled (^{13}C or ^{15}N) amino acids (Arg and Lys), which are incorporated into proteins during biosynthesis *in vivo*. SILAC enables robust MS quantification of up to three multiplexed conditions, which can be quantitatively compared within the same MS1 spectra. Thus, samples can already be combined early on when cells are still intact, which reduces technical variability compared to LFQ or chemical labeling-based quantification strategies. However, SILAC is restricted to only three samples that can be quantitatively compared since isotopically-labeled Arg and Lys only give rise to three distinct SILAC media (“light”, “medium” and “heavy”). Moreover, SILAC results in reduced MS sequencing depth due to more complex MS1 spectra.

1.19 Aim of this study

R-loops are considered a double-edged sword for cells. On the one hand, they participate in the regulation of various cellular pathways such as transcription, chromosome segregation during mitosis, and DNA repair. On the other hand, R-loops can cause DNA damage through

transcription-replication conflicts or processing of the nucleotide excision repair machinery. Thus, R-loop abundance needs to be tightly balanced to prevent R-loop-dependent genomic instability, while maintaining the regulatory functions of R-loops. Cells have developed dedicated protein machinery to oppose excessive R-loop formation. Mutations within R-loop regulating genes are associated with neurodegenerative diseases and cancer, in particular MDS and AML. Since the whole spectrum of R-loop regulators in human cells remains poorly characterized, we aim to decipher the R-loop proximal proteome in human cells. To this end, we plan to establish a proximity labeling approach, by targeting the ascorbate peroxidase APEX2 to R-loops through conjugation to the hybrid binding domain of RNaseH1. We want to dissect the R-loop proximal proteome in functional clusters to understand which pathways and groups of proteins associate with R-loops on native chromatin. Furthermore, we will functionally characterize a subset of the identified R-loop proximal proteins in human cells to understand if and how they contribute to maintaining R-loop homeostasis and preventing genomic instability. We want to investigate how dysregulation of R-loops in the presence of pathogenic variants of R-loop regulators drives genomic instability and contributes to the development of MDS/AML.

2 Results

2.1 RDProx allows the identification of R-loop proximal proteins

Scheduled R-loops are beneficial for cells since they contribute to the regulation of crucial molecular mechanisms, such as transcription and DNA repair. However, excessive R-loop formation can lead to replication stress, DNA damage, and ultimately genomic instability. Therefore, cells rely on tight regulation of R-loop abundance on chromatin to prevent R-loop-dependent genomic instability. Augmented R-loop levels have been linked to neurodevelopmental diseases, myelodysplastic syndromes, and cancer. To gain further insight into which protein networks ensure R-loop homeostasis in human cells, we established a workflow called RNA-DNA proximity proteomics (RDProx) to identify R-loop proximal proteins in human cells under native conditions. RDProx relies on the RNA-DNA hybrid binding function of the hybrid binding domain (HBD) of RNaseH1. As a negative control, we introduced three point mutations in the HBD (HBD-WKK), which abrogate the binding affinity towards RNA-DNA hybrids (Figure 6a). Both domains were conjugated to the promiscuous soybean ascorbate peroxidase (APEX2)[320], which allows biotinylation of proximal proteins in the presence of biotin-phenol and H_2O_2 (Figure 6a). To test

the binding affinity of HBD and HBD-WKK to RNA-DNA hybrids *in vitro*, we purified both domains and probed them with five different 6-FAM-conjugated nucleic acid substrates. Electron mobility shift assay (EMSA) revealed exclusively strong binding of the HBD to RNA-DNA hybrids, while HBD-WKK displayed dramatically decreased affinity (Figure 6b). Accordingly, the HBD associated with RNA-DNA hybrids *in vitro* with a k_d = 190 nm +/- 30 nm, whereas the HBD-WKK bound with a $K_{1/2}$ = 7.5 µM +/- 1.3 µM (Figure 6c). Moreover, only GFP-tagged HBD and not HBD-WKK was retained on chromatin under pre-extraction conditions (Figure 6d).

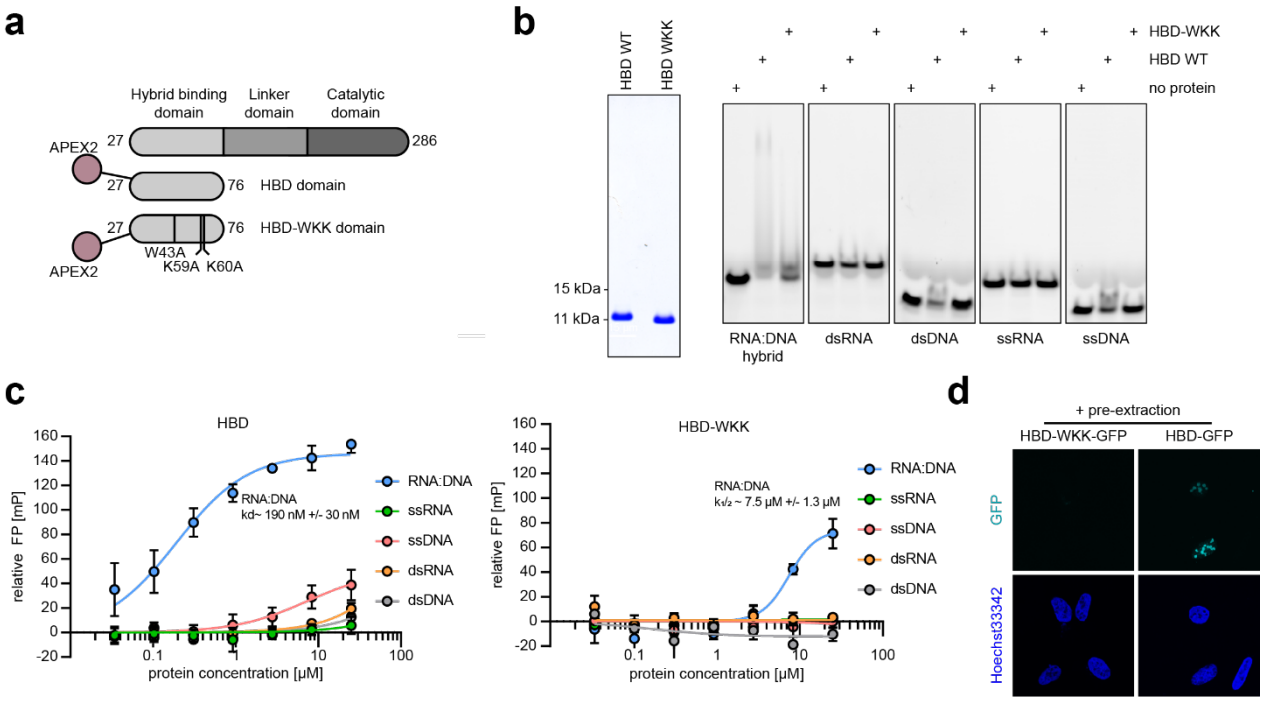


Figure 6: Hybrid binding domain of RNaseH1 possesses a strong binding affinity towards RNA-DNA hybrids in vitro

a: Schematic representation of M27-RNaseH1 domains. The wild-type version of the hybrid binding domain (HBD) of RNaseH1 was cloned as well as the mutant with indicated WKK point mutations to abrogate the binding function. HBD and HBD-WKK were fused to the ascorbate peroxidase APEX2 on the N-terminus. Numbers indicate amino acid positions. **b:** Electro Mobility Shift Assay (EMSA) using 25 µM purified HBD or HBW-WKK mutant and indicated 6-FAM-conjugated oligonucleotides. Protein inputs are shown left. Data derived from n = 1 experiment **c:** Fluorescence polarization (FP) assay of wild-type HBD and HBD-WKK mutant. Titrated domains were incubated together with indicated 20 nM 6-FAM-conjugated oligonucleotide substrates. Data of n = 3 independent experiments with individually thawed proteins are represented as mean +/- standard deviation. Colored lines represent either a Michalis-Menten or a sigmoidal fit. The resulting k_d or $k_{1/2}$ values for the association to the RNA-DNA hybrid substrate are displayed. **d:** Immunofluorescence analysis of the retention of GFP-tagged HBD or HBD-WKK on chromatin after pre-extraction with 0.4% NP-40 for 30 min on ice. DNA was counterstained with Hoechst33342 (blue), and GFP signal (cyan). Representative images of n = 2 biologically independent experiments.

We transiently expressed APEX2-HBD and APEX2-HBD-WKK fusion proteins in U2OS cells to test the expression efficiency and activity of the peroxidase. Efficient biotinylation was apparent specifically after the addition of biotin-phenol and H_2O_2, whereas only endogenous biotinylation was observed in the absence of H_2O_2 (Figure 7a). RDProx was combined with quantitative mass spectrometry (MS) to map the R-loop proximal proteome in HEK293T cells. Stable isotope labeling with amino acids in cell culture (SILAC) enabled quantitative comparisons between proteins identified in the APEX2-HBD (heavy) condition and the APEX2-HBD-WKK negative control (light) condition (Figure 7b). Proteins which were proximal to the bait proteins were biotinylated, enriched using NeutrAvidin beads, and analyzed by liquid chromatography (LC)-tandem MS/MS (Figure 7b). 612 R-loop proximal proteins were identified by RDProx in three biologically independent experiments with high reproducibility ($r > 0.85$) (Figure 7c, 7d). We further categorized the identified proteins into two confidence intervals. The high confidence interval Tier1 consists of 312 proteins, which were identified with a $\log_2$ fold change (FC) > 2 and a false discovery rate (FDR) < 0.01. 300 proteins were mapped within Tier2 with a $\log_2$FC > 1 and a FDR < 0.01. Among proteins identified in Tier1 were well-characterized R-loop regulators such as TOP1, the RNA helicase AQR, components of the exosome complex (EXOSC7, EXOSC10), the THO complex (THOC1/2/6, THOC6, ALYREF), and the RNA-DNA helicase DDX39B (Figure 7d). Moreover, we identified subunits of the RNA polymerase II (POLR2A), PARP1, and transcription termination factor XRN2. RnaseH2A, the single-stranded DNA binding proteins RPA1/2, TOP2A/B were identified in Tier2 (Figure 7d).

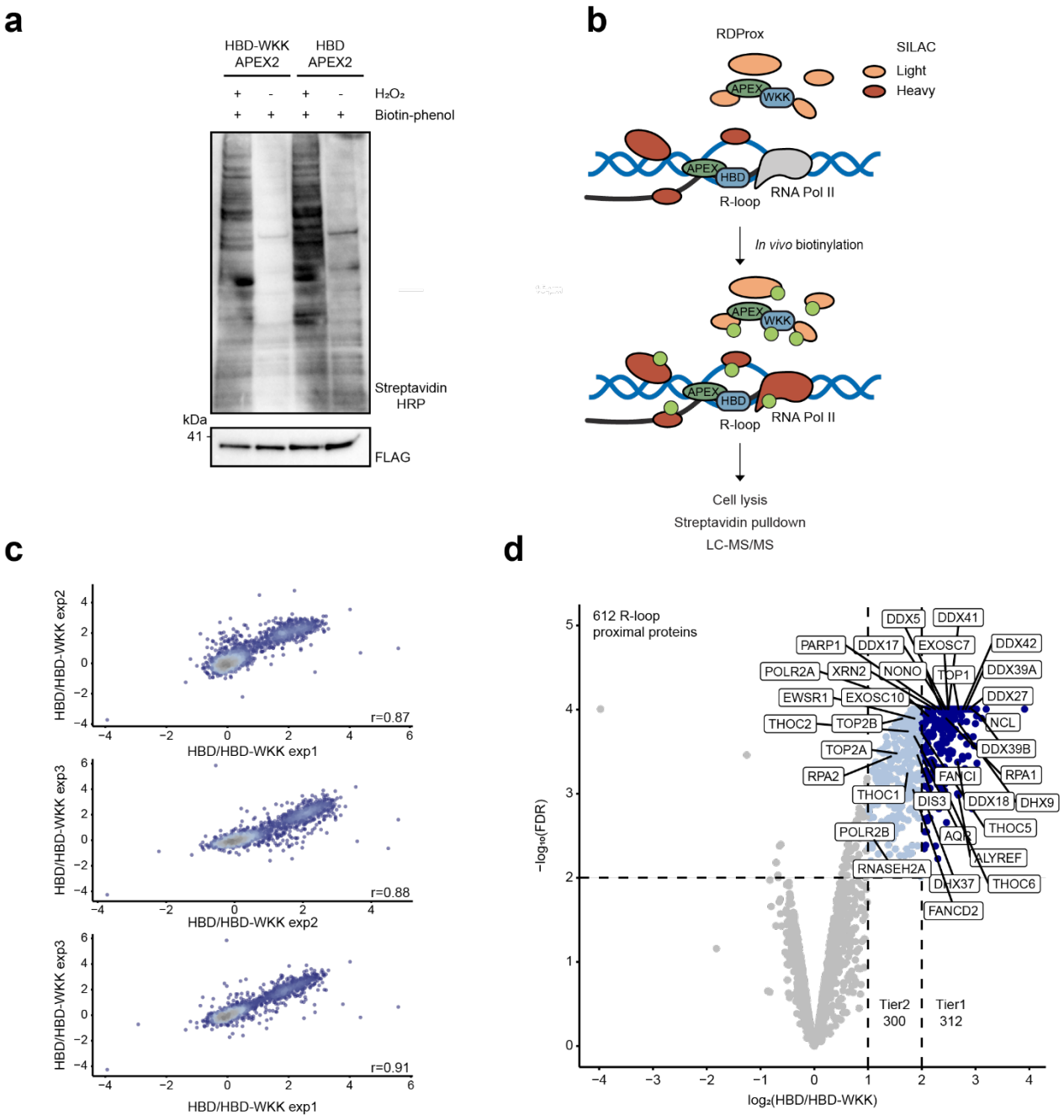

a
HBD-WKK APEX2
HBD APEX2
+ - + - H2O2
+ + + + Biotin-phenol
Streptavidin HRP
kDa
41
FLAG
b
RDProx
SILAC
Light
Heavy
APEX
WKK
HBD
R-loop
RNA Pol II
In vivo biotinylation
Cell lysis
Streptavidin pulldown
LC-MS/MS
c
HBD/HBD-WKK exp2
HBD/HBD-WKK exp1
r=0.87
HBD/HBD-WKK exp3
HBD/HBD-WKK exp2
r=0.88
HBD/HBD-WKK exp3
HBD/HBD-WKK exp1
r=0.91
d
612 R-loop proximal proteins
DDX5
DDX41
PARP1
DDX17
EXOSC7
DDX42
POLR2A
XRN2
NONO
TOP1
DDX39A
EWSR1
EXOSC10
DDX27
THOC2
TOP2B
NCL
TOP2A
DDX39B
RPA2
FANCI
RPA1
THOC1
DIS3
DDX18
DHX9
POLR2B
THOC5
AQR
ALYREF
RNASEH2A
DHX37
THOC6
FANCD2
Tier2 300
Tier1 312
−log10(FDR)
log2(HBD/HBD-WKK)

Figure 7: Mapping R-loop-proximal proteome on native chromatin by RDProx

a: Representative western blot of biotinylated proteins in U2OS whole-cell extracts after HBD-APEX2 or HBD-WKK-APEX2 proximity labeling from n = 2 biologically individual experiments. **b:** Schematic representation of the RDProx workflow for identification of R-loop-proximal proteins. HBD or HBD-WKK fused N-terminally to APEX2 were transiently expressed in light or heavy SILAC-labeled HEK293T cells. Biotinylation was induced upon the addition of 500 µM biotin-phenol for 2 h at 37°C and 1 mM H_2O_2 for 2 min at room temperature. Samples were pooled after cell lysis and biotinylated proteins were purified using NeutrAvidin beads. Denatured proteins were separated by SDS-PAGE and in-gel digested before LC-MS/MS analysis. **c:** Multi-scatter plots indicating the correlation between the n = 3 biologically independent RDProx SILAC experiments in HEK293T cells. Pearson correlation is indicated below each plot. Coloring indicates the density of the population (blue=less dense, grey= high density). **d:** Volcano plot of protein groups identified by RDProx in n = 3 biologically independent experiments. Mean $\log_2$ ratios of all replicates between HBD and HBD-WKK are plotted against the $-\log_{10}$ FDR. The FDR and enrichment were calculated using Limma103. Significantly enriched proteins are highlighted in blue (FDR < 0.01). Light blue indicates proteins in Tier2 (300 proteins) above a 2-fold change of the mean ratio and dark blue indicates proteins in Tier1 (312 proteins) with a 4-fold change or higher.

Different protein clusters were evident within the R-loop proximal proteome, indicating functional protein interaction networks and suggesting specific molecular pathways to play a role at R-loops. Enriched were protein clusters involved in transcriptional regulation, chromatin organization, mRNA export, splicing regulation, m^6A regulation, cell cycle regulation, DNA replication, and repair (Figure 8a). Within these clusters, proteins were enriched with domains required for nucleic acid binding and processing such as RRM, DEAD/DEAH, CID domain, RNA polymerase II binding domain, SAP domain, MCM OB domain, and K Homology domain (Figure 8b).

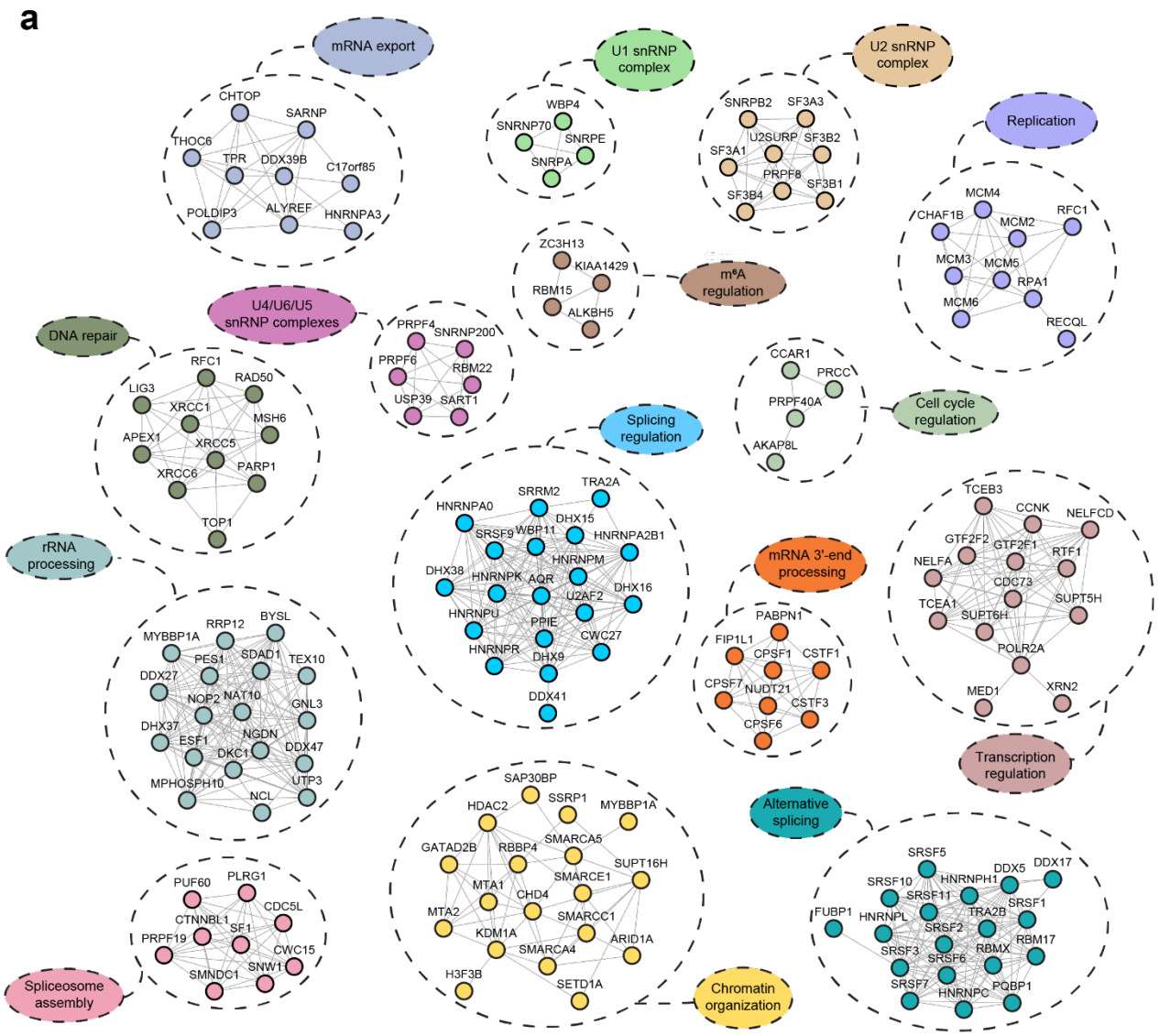

a
mRNA export
CHTOP
SARNP
THOC6
TPR
DDX39B
C17orf85
POLDIP3
ALYREF
HNRNPA3
U1 snRNP complex
WBP4
SNRNP70
SNRPE
SNRPA
U2 snRNP complex
SNRPB2
SF3A3
U2SURP
SF3B2
SF3A1
PRPF8
SF3B1
SF3B4
Replication
MCM4
CHAF1B
RFC1
MCM2
MCM3
MCM5
RPA1
MCM6
RECQL
ZC3H13
KIAA1429
RBM15
ALKBH5
m⁶A regulation
U4/U6/U5 snRNP complexes
PRPF4
SNRNP200
PRPF6
RBM22
USP39
SART1
DNA repair
RFC1
RAD50
LIG3
XRCC1
MSH6
APEX1
XRCC5
XRCC6
PARP1
TOP1
CCAR1
PRCC
PRPF40A
AKAP8L
Cell cycle regulation
Splicing regulation
SRRM2
TRA2A
HNRNPA0
DHX15
SRSF9
WBP11
HNRNPA2B1
HNRNPM
DHX38
HNRNPK
AQR
U2AF2
DHX16
HNRNPU
PPIE
CWC27
HNRNPR
DHX9
DDX41
rRNA processing
RRP12
BYSL
MYBBP1A
PES1
SDAD1
TEX10
DDX27
NOP2
NAT10
GNL3
DHX37
ESF1
NGDN
DDX47
DKC1
MPHOSPH10
UTP3
NCL
mRNA 3'-end processing
PABPN1
FIP1L1
CPSF1
CSTF1
CPSF7
NUDT21
CSTF3
CPSF6
TCEB3
CCNK
NELFCD
GTF2F2
GTF2F1
RTF1
NELFA
CDC73
TCEA1
SUPT6H
SUPT5H
POLR2A
MED1
XRN2
Transcription regulation
Alternative splicing
SRSF5
DDX5
DDX17
SRSF10
HNRNPH1
SRSF11
SRSF1
FUBP1
HNRNPL
TRA2B
SRSF2
RBM17
SRSF3
SRSF6
RBMX
SRSF7
HNRNPC
PQBP1
SAP30BP
SSRP1
HDAC2
MYBBP1A
SMARCA5
GATAD2B
RBBP4
SUPT16H
SMARCE1
MTA1
CHD4
MTA2
SMARCC1
KDM1A
ARID1A
SMARCA4
H3F3B
SETD1A
Chromatin organization
Spliceosome assembly
PUF60
PLRG1
CDC5L
CTNNBL1
SF1
PRPF19
CWC15
SNW1
SMNDC1
b
RRM domain
3.88e−57
Helicase, C−terminal
6.5e−17
Helicase superfamily 1/2 ATP−binding domain
6.61e−17
DEAD/DEAH box helicase domain
5.84e−15
CID domain
8.09e−09
MCM N−terminal domain
9.94e−08
RNA polymerase II binding domain
3e−07
SAP domain
5.92e−07
MCM OB domain
6.33e−07
K Homology domain
7.65e−07
0
10
20
30
40
50
−log₁₀(adjusted P−value)

Figure 8: RDProx identifies specific protein networks in the proximity of R-loops

a: Functional interaction network of proteins identified by RDProx. Genes were manually annotated based on literature and corresponding GO terms (Biological Process and Molecular Function). Clusters were generated based on manual annotation. Edges between the nodes indicate interactions based on STRING with a confidence score equal to or above 0.7. **b:** PFAM InterPro Domain enrichment analysis using EnrichR. Adjusted p-values were derived using a two-sided Fisher´s exact test with Bonferroni correction.

Before this study, two proteomic screens to identify R-loop interacting proteins have been published. One screen employed the S9.6 antibody, which possesses a high affinity towards RNA-DNA hybrids, to co-immunoprecipitate nuclear RNA-DNA hybrids and associated proteins[152]. The other screen probed either RNA-DNA hybrid substrates or dsDNA substrates with nuclear lysates to enrich proteins with higher affinity to RNA-DNA hybrids than dsDNA[214]. For unbiased comparison between RDProx and the two screens, we took the significantly enriched proteins from each screen (for RDProx this amounted to 311 proteins, for S9.6 IP to 469 proteins, and the in vitro pulldown to 364 proteins) and performed gene set enrichment analysis (GSEA) based on GO-terms for biological processes (BP). The most significant terms for exclusively identified proteins in the S9.6-based screen were “rRNA processing” and “ribosome biogenesis”, while “rRNA metabolic process” and “ribosome biogenesis” were the most significantly enriched terms in the *in vitro* RNA-DNA hybrid pulldown screen (Figure 9a, 9b). This was not the case for proteins exclusively identified by RDProx, for which the most significantly enriched terms were “mRNA processing” and “mRNA splicing” (Figure 9a, 9b).

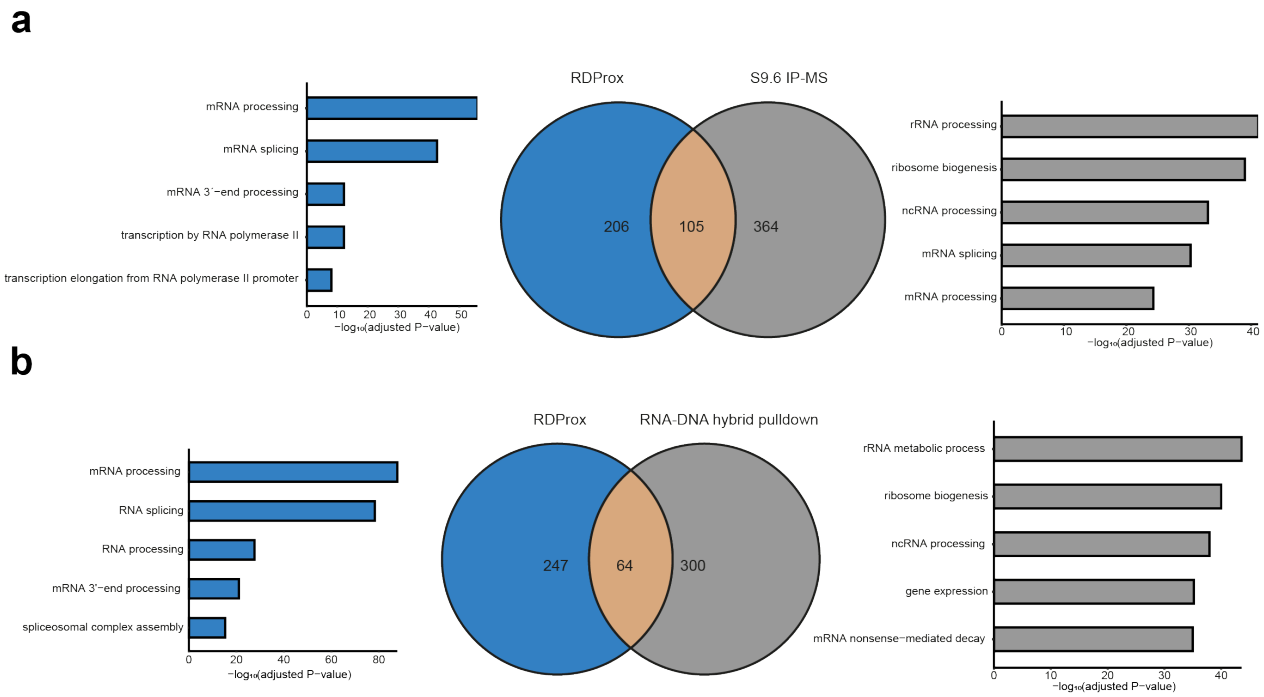


Figure 9: Overlap between RDProx and previous R-loop proteomics screens

a: GO Biological Process analysis of protein groups identified in S9.6 immunoprecipitation-based mass spectrometry screen (Cristini et al., 2018) and RDProx Tier 1[152]. The 5 GO-BP terms with the lowest adjusted p-value (two-sided Fisher´s exact test, Bonferroni correction) are shown. **b:** GO Biological Process analysis of protein groups identified in RNA-DNA hybrid in vitro pulldown mass spectrometry screen (Wang et al., 2018) and RDProx Tier 1[214]. The 5 GO-BP terms with the lowest adjusted p-value (two-sided Fisher´s exact test, Bonferroni correction) are shown.

2.2 DDX41 prevents spontaneous DNA damage and replication stress

Proteins containing DEAD/DEAH box domains were enriched in the R-loop proximal protein network mapped by RDProx. A few DEAD-box helicases such as DDX5 and DDX19 have recently been described as R-loop regulators[216,217,353]. Loss of R-loop homeostasis frequently leads to R-loop-dependent genomic DNA damage. To assess whether there were also other DEAD/DEAH-box helicases regulating R-loop levels in human cells, we monitored Ser139 phosphorylation on the histone variant H2AX (yH2AX) as a proxy for spontaneous DNA damage upon knockdown of DDX27, DDX41, DDX42, DHX37, and DDX39A. Depletion of the known R-loop regulator AQR served as a positive control since the loss of AQR has been reported to lead to spontaneous DSB formation[207]. Significant yH2AX formation was only apparent after knockdown of the positive control AQR and DDX41 (Figure 10a, 10b). DDX41 is a poorly characterized DEAD-box helicase, which is frequently mutated in familial cases of MDS and AML[263,270]. Since augmented R-loop formation has been proposed as a unifying mechanism for MDS and AML development, we further characterized a potential role of DDX41 in R-loop regulation and tumor suppression[210]. In addition

to increased γH2AX levels, DDX41 knockdown cells displayed significantly more 53BP1 foci on chromatin, which were suppressed by overexpression of nuclear RNaseH1 under the control of a doxycycline-inducible promoter (Figure 10c). Moreover, spontaneous γH2AX formation in the absence of DDX41 was partially rescued by overexpression of nuclear RNaseH1, pointing towards R-loop-dependent genomic instability (Figure 10d). Accordingly, spontaneous DSB formation after DDX41 knockdown, measured by neutral comet assay, was rescued by overexpression of nuclear RNaseH1 (Figure 10e).

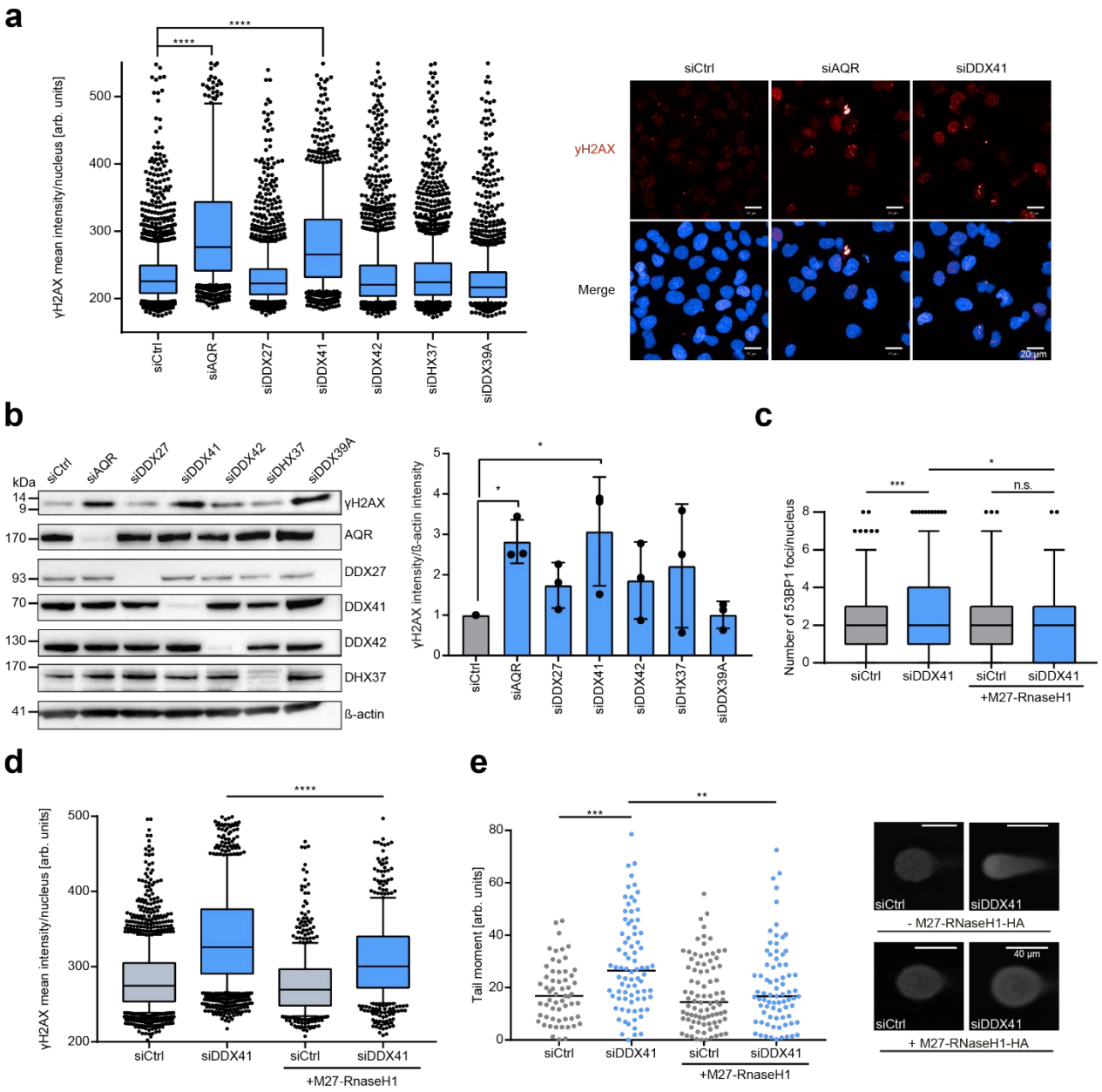


Figure 10: DDX41 knockdown cells experience spontaneous R-loop-dependent DNA damage

a: Immunofluorescence analysis of yH2AX in U2OS cells 48 h after indicated knockdowns. Centerlines of boxplots indicate the median, the limits the 25th-75th percentile, whiskers the 10th-90th percentile, dots outliers. Representative data of n = 3 biologically independent experiments; p-values ($p < 0.0001$, $p = 0.1788$, $p < 0.0001$, $p = 0.8259$, $p > 0.9999$, $p = 0.9569$) were derived from >1000 cells using one-way ANOVA with Tukey correction for multiple comparisons. Representative images of yH2AX (red) staining and Hoechst33342 (blue). Scale bars - 20 µm. **b:** (left) Quantification of the ratio between γH2AX intensity and ß-actin. Data of n = 3 biologically independent experiments are represented as the mean +/- standard deviation. P-values ($p = 0.0297$, $p = 0.2627$, $p = 0.0177$, $p = 0.2627$, $p = 0.1876$, $p = 0.8433$) derived using Kruskal-Wallis test with Benjamini-Hochberg correction for multiple comparisons. (Right) Representative Western blot of yH2AX in U2OS cells after the indicated knockdown. Bar plot shows qPCR after DDX39A knockdown. Data are represented as the mean +/- standard deviation of n = 2 biologically independent experiments. **c:** Immunofluorescence analysis of 53BP1 foci 48 h after indicated knockdowns in U2OS cells +/- the expression of M27-RNaseH1-GFP. Boxplot limits indicate the 25th to 75th percentile, whiskers the 5th-95th percentile, and the center line the median. Dots represent outliers. Three outlier data points above 8 foci/cell are cut off. Representative data of n = 2 biologically independent experiments. P-values ($p = 0.008$, $p = 0.0455$, $p = 0.3005$) derived from $n > 170$ cells using one-way-ANOVA with Tukey correction for multiple comparisons. **d:** Immunofluorescence analysis of yH2AX in U2OS cells +/- doxycycline-inducible GFP-tagged M27-RNaseH1. Quantification of cells with medium GFP intensity (medium M27-RNaseH1 expression). Representative box plot of n = 2 biologically independent experiments. The Center of boxplots indicates the median, limits the 25th-75th percentile, whiskers the 10th-90th percentile, dots outliers. $P < 0.0001$ derived from $n > 500$ cells using a two-sided Mann-Whitney test. **e:** Single-cell electrophoresis of U2OS cells 48 h after knockdown +/- doxycycline-inducible expression of HA-tagged M27-RNaseH1. Representative images are displayed (right). Scale bars - 40 µm. Dots depict individual tail moments, black lines the median. Representative results from n = 2 biologically independent experiments. P-values ($p = 0.0001$, $p = 0.0031$) were derived from $n > 50$ cells using one-way ANOVA with Tukey correction for multiple comparisons.

R-loop-dependent genomic instability frequently arises through transcription-replication conflicts, thus leading to S-phase-specific DNA damage. To assess whether DDX41 knockdown induces spontaneous DNA damage specifically during S-phase, we measured yH2AX intensity in different stages of the cell cycle. We observed significantly higher yH2AX levels during all phases of the cell cycle (G1, S, and G2/M) (Figure 11a). However, the strongest yH2AX increase was apparent during the S-phase of the cell cycle, suggesting that DDX41 knockdown cells suffer from replication stress (Figure 11a). Indeed, Ser33 phosphorylation of RPA, a surrogate marker for replication stress, was significantly higher after knockdown of AQR and DDX41, but not DDX27, DDX41, DHX37, or DDX39A (Figure 11b). Moreover, replication fork progression was reduced in DDX41 knockdown cells to a similar extent as in cells treated with a low dose (100 nM) of the DNA polymerase inhibitor aphidicolin (Figure 11c). Phosphorylation of Ser33 on RPA is a proxy for the activation of the main replication stress response kinase ATR. Accordingly, treatment of DDX41 knockdown cells with the ATR inhibitor VE-821 significantly reduced Ser33 phosphorylation on RPA, indicating that ATR is activated in the absence of DDX41 (Figure 11d). Moreover, either depletion of DDX41 or overexpression of pathogenic AML DDX41 variants

(L237F/P238T and R525H) led to sensitization of U2OS and OCI-AML3 cells, respectively (Figure 11e). Sensitivity of AML cells expressing pathogenic DDX41 mutants to ATR inhibition suggests that these cells also experience replication stress and ATR activity.

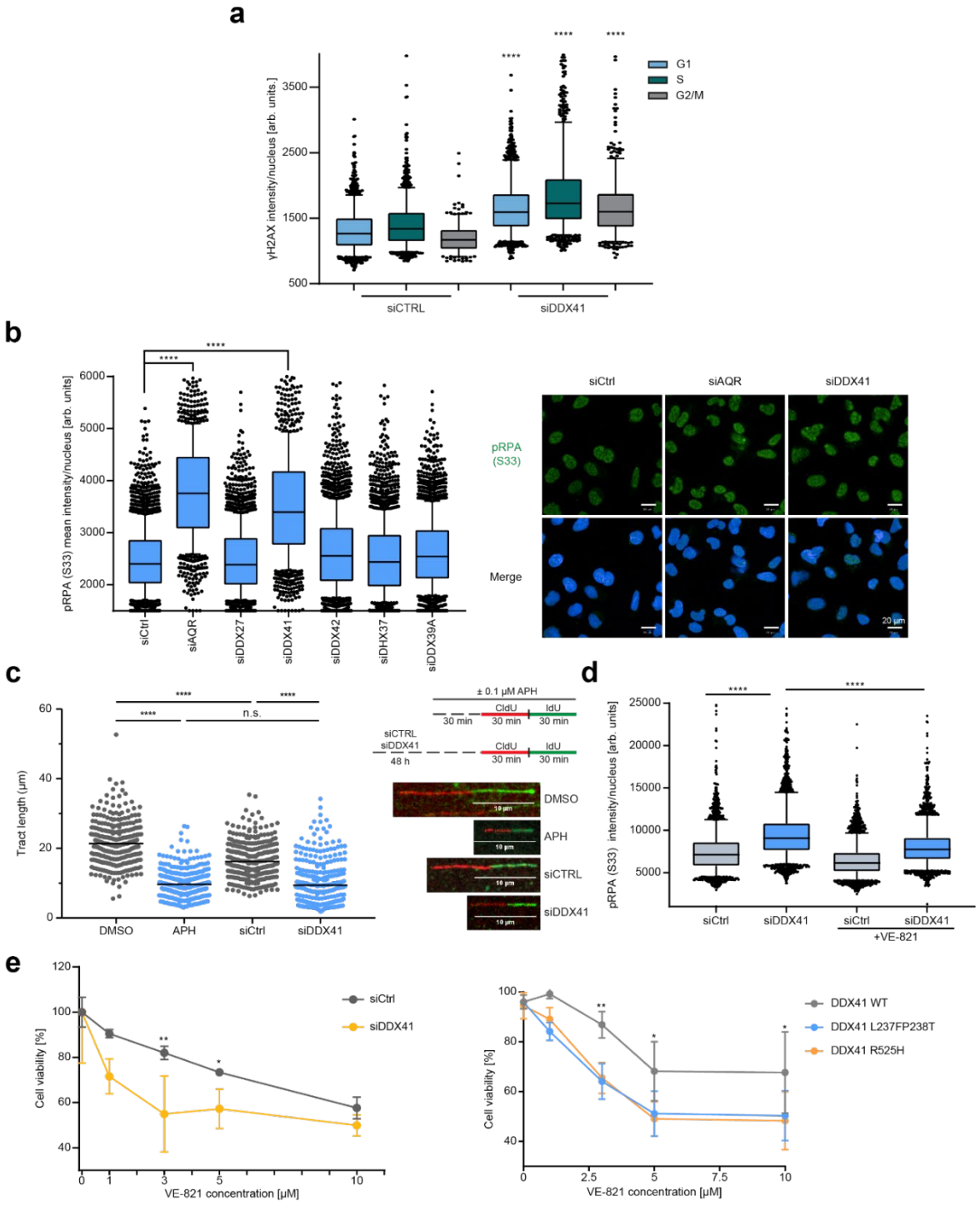


Figure 11: Loss of DDX41 results in increased replication stress and dependency on ATR

a: Immunofluorescence analysis of γH2AX after 48 h knockdown with the indicated siRNAs in different cell cycle stages: G1 (blue), S (green), G2/M (grey). Center-line of box plots represents the median, whiskers indicate the 10th-90th percentile, and box limits the 25th-75th percentile. Dots represent outliers. Representative data of n = 2 biologically independent experiments. P-values ($p < 0.0001$, $p < 0.0001$, $p < 0.0001$) derived by one-way-ANOVA with Tukey correction for multiple comparisons. **b:** Immunofluorescence analysis of pRPA (Ser33) in U2OS 48 h after indicated knockdowns. Representative images (right): Hoechst33342 (blue), pRPA (Ser33) (green). The Center of boxplots indicates the median, limits the 25th-75th percentile, whiskers the 10th-90th percentile, and dots outliers. Representative data of n = 3 biologically independent experiments are displayed. P-values ($p < 0.0001$, $p > 0.9999$, $p < 0.0001$, $p < 0.0001$, $p = 0.9384$, $p < 0.0001$) were derived from n > 1000 cells using one-way ANOVA with Tukey correction for multiple comparisons. Scale bars - 20 µm. **c:** DNA fiber spreading assay of U2OS cells after 48 h knockdown of DDX41. Controls were either treated with DMSO or 100 nM aphidicolin for 1.5 h. Representative images (white line indicates 10 µm scale) and quantifications of fiber tract length. Dots represent individual values and the black line the median. At least 260 fibers were quantified across n = 1 experiment. P-values ($p < 0.0001$, $p < 0.0001$, $p < 0.0001$, $p = 0.5794$) were derived using one-way ANOVA with Tukey correction for multiple comparisons. **d:** Immunofluorescence analysis of pRPA (Ser33) in U2OS cells after indicated knockdowns for 48 h and 6 h treatment with DMSO or 10 µM VE-821. Boxplot limits indicate the 25th to 75th percentile, whiskers the 5th-95th percentile, and the center line the median. Outliers are represented by dots. P-values ($p < 0.0001$, $p < 0.0001$) derived by one-way-ANOVA with Tukey correction for multiple comparisons. N = 2 **e:** (left) Cell titer blue assay in U2OS cells after indicated knockdowns. Data from n = 3 biologically independent experiments is represented as mean +/- standard deviation. P-values ($p = 0.033$, $p = 0.0038$) derived using two-way ANOVA with Tukey correction for multiple comparisons. (Right) Cell titer blue assay in OCI-AML3 cells after expression of DDX41 WT, R525H, or L237FP238T mutants. Data from n = 3 biologically independent experiments is represented as mean +/- standard deviation. P-values ($p = 0.0044$, $p = 0.0078$, $p = 0.0374$, $p = 0.0179$, $p = 0.338$, $p = 0.0162$) derived using two-way ANOVA with Tukey correction for multiple comparisons.

2.3 DDX41 unwinds RNA-DNA hybrids *in vitro* and regulates R-loop levels *in vivo*

Since DDX41 knockdown cells experienced replication stress and predominantly R-loop-dependent DNA damage during S-phase, we explored a potential R-loop regulating role of DDX41. RDProx MS analysis identified DDX41 to be proximal to R-loops. We confirmed the proximity of DDX41 to R-loops by coupling RDProx with Western blot analysis (Figure 12a). Moreover, the proximity of DDX41 was R-loop-dependent, since overexpression of nuclear RNaseH1 decreased RDProx signal (Figure 12a). Proximity of DDX41 to R-loops was further validated by Proximity ligation assay (PLA) of endogenous DDX41 and exogenous GFP-tagged HBD (Figure 12b). The Association of DDX41 to GFP-HBD was significantly higher than that of the non-binding GFP-HBD-WKK mutant, suggesting that the PLA signal specifically resulted from the proximity of DDX41 to RNA-DNA hybrids (Figure 12c). To assess whether DDX41 was not only proximal to R-loops but also regulates global R-loop levels, we performed dot blot analysis using the S9.6 antibody. U2OS cells displayed globally increased R-loop levels after knockdown of AQR and DDX41 (Figure 12d). R-loop accumulation after loss of DDX41 was suppressed by expression of nuclear RNaseH1 (Figure 12d). Importantly, *in vitro* digestion with bacterial RNase H depleted the

S9.6 signal, confirming the specificity of the antibody to RNA-DNA hybrids (Figure 12d). Decreasing trust in the S9.6 antibody in the field, due to unspecific RNA binding prompted us to use an antibody-independent approach to confirm the dot-blot results[150]. Retention of GFP-tagged HBD - expressed under a doxycycline-inducible promoter - on chromatin under pre-extraction conditions served as a proxy for R-loop levels in U2OS cells. Knockdown of DDX41 and the positive control AQR led to significantly increased R-loop levels based on GFP-HBD chromatin retention (Figure 12e). Transcriptional inhibition using DRB suppressed R-loop accumulation in the absence of DDX41, validating the sensitivity of the approach towards R-loops (Figure 12e).

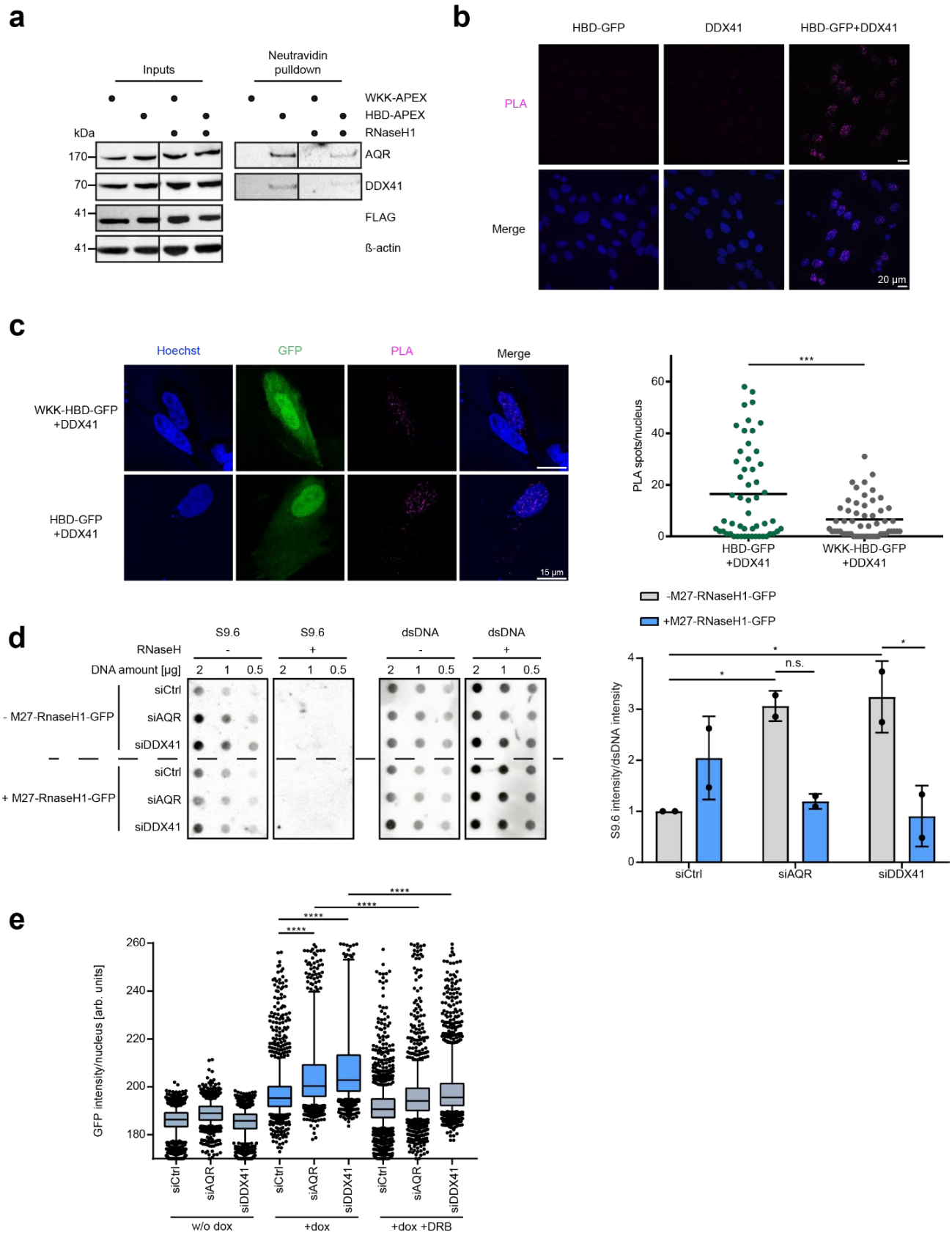

a
Inputs
Neutravidin pulldown
WKK-APEX
HBD-APEX
RNaseH1
kDa
170
70
41
41
AQR
DDX41
FLAG
ß-actin
b
HBD-GFP
DDX41
HBD-GFP+DDX41
PLA
Merge
20 µm
c
Hoechst
GFP
PLA
Merge
WKK-HBD-GFP +DDX41
HBD-GFP +DDX41
15 µm
***
PLA spots/nucleus
HBD-GFP +DDX41
WKK-HBD-GFP +DDX41
d
S9.6
S9.6
dsDNA
dsDNA
RNaseH
DNA amount [µg]
siCtrl
siAQR
siDDX41
- M27-RnaseH1-GFP
+ M27-RnaseH1-GFP
-M27-RNaseH1-GFP
+M27-RNaseH1-GFP
n.s.
S9.6 intensity/dsDNA intensity
e
GFP intensity/nucleus [arb. units]
w/o dox
+dox
+dox +DRB

Figure 12: DDX41 associates with and regulates R-loops *in vivo*

a: RDProx-Western blot analysis AQR and DDX41 in U2OS cells +/- expression of M27-RNaseH1. Data from n = 1 experiment. **b:** Proximity ligation assay (PLA) between endogenous DDX41 and GFP-tagged HBD or HBD-WKK. Representative images of nuclear PLA spots. **c:** (left) Representative images of Proximity Ligation assay (PLA) between endogenous DDX41 and exogenously expressed HBD-GFP or HBD-WKK-GFP. Data derived from n = 1 individual experiment. The scale bar represents 15 µm. (right) Quantification of nuclear PLA spots. Dots represent results from individual cells, Black line indicates the median. P-value < 0.0001 derived from > 50 cells from n = 1 experiment using two-sided Student´s t-test. **d:** S9.6 and double-stranded DNA dot blot analysis of U2OS cells expressing GFP-tagged M27-RNaseH1 upon doxycycline (dox) after 48 h of indicated knockdowns. Representative images (left). Data are represented as mean +/- standard deviation (right). Black dots represent individual results from n = 2 biologically independent experiments. P-values (p = 0.437, p = 0.653, p = 0.338, p = 0.0281) derived using one-way ANOVA with Tukey correction for multiple comparisons. **e:** HBD-GFP retention assay after indicated 48 h knockdowns in U2OS cells. Control cells were treated with 100 µM DRB for 3 h. The Center of boxplots indicates the median of the population, limits the 25th-75th percentile, whiskers the 10th-90th percentile and dots represent outliers. Representative data of n = 3 biologically independent experiments are shown. P-values (p < 0.0001, p < 0.0001, p < 0.0001, p < 0.0001) derived from n > 1000 cells using one-way ANOVA with Tukey correction for multiple comparisons.

The finding that global R-loop levels were increased in DDX41 knockdown cells prompted us to investigate whether DDX41 regulates R-loop levels by unwinding RNA-DNA hybrids. To this end, we purified full-length DDX41, the pathogenic R525H AML mutant, and a short version of DDX41 (153-410), which lacks the helicase domain (Figure 13a, 13b). *In vitro*, recombinant DDX41 preferentially associated with RNA-DNA hybrids (k_D=2.5µM +/- 1.4 µM) compared to four other nucleic acid substrates (Figure 13c). Moreover, incubation of DDX41 with an RNA-DNA hybrid substrate with an ssDNA overhang led to the hydrolysis of ATP to ADP in a DDX41 concentration-dependent manner (Figure 13d). To test the potential RNA-DNA hybrid unwinding activity of DDX41, we established a FRET-based unwinding assay that is based on the increase of fluorescence intensity, when an RNA-DNA hybrid substrate containing a fluorophore-conjugated DNA and a quencher-conjugated RNA are displaced. DDX41 not only hydrolyzed ATP but also unwound the RNA-DNA hybrid substrate in a concentration-dependent manner (Figure 13e). We observed homeostasis between unwinding and re-annealing of the substrate in the presence of DDX41, which was shifted towards unwinding (increased fluorescence intensity due to displacement of the fluorophore-labeled DNA) upon addition of unlabeled DNA (Figure 13e). Importantly, the DDX41 mutant lacking the helicase domain was unable to unwind RNA-DNA hybrids (Figure 13f). Furthermore, the DDX41 mutant containing the pathogenic R525H mutation within the helicase domain displayed significantly reduced unwinding activity (Figure 13f).

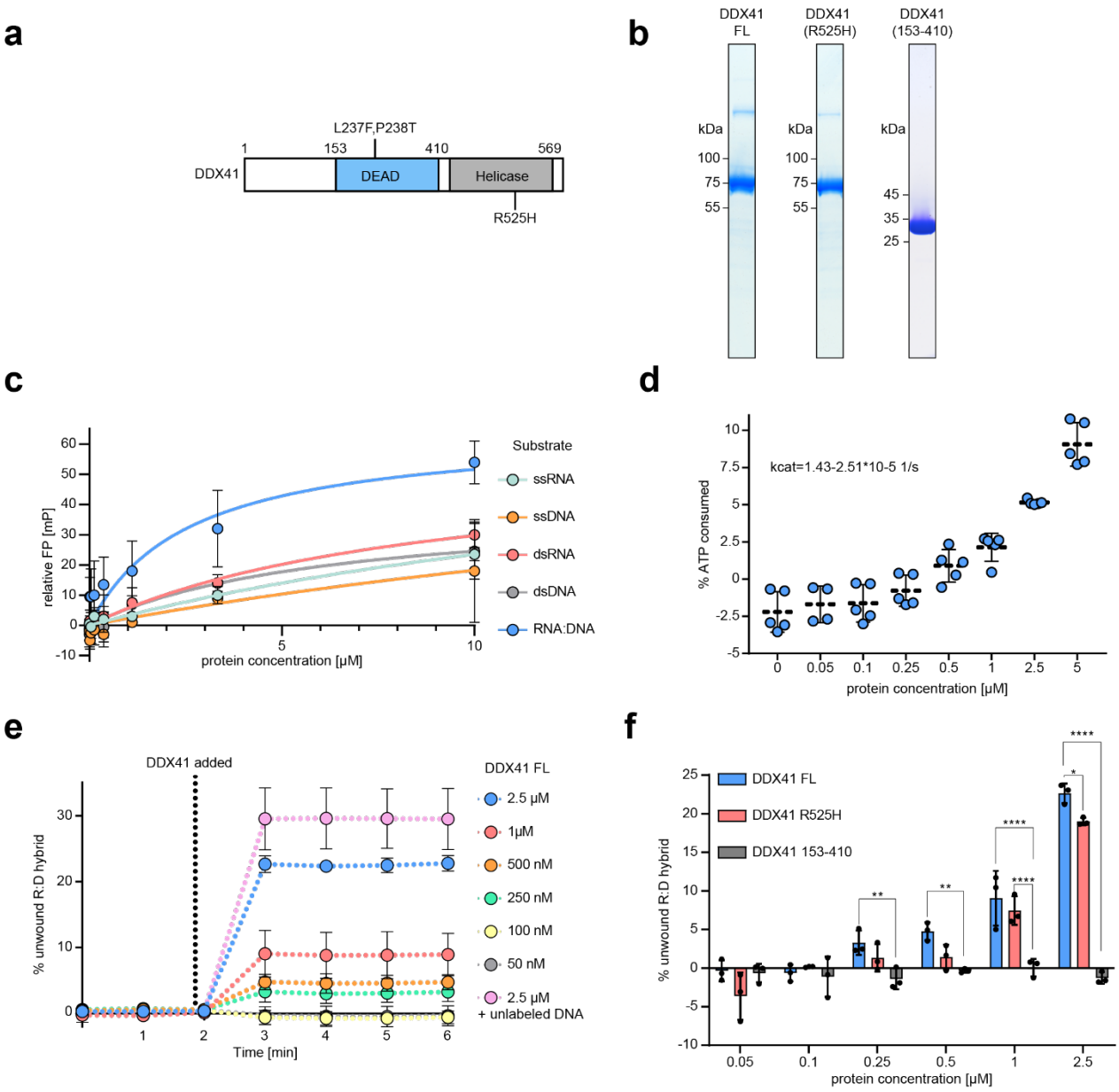


Figure 13: DDX41 binds and unwinds RNA-DNA hybrids *in vitro*

a: Schematic representation of full-length DDX41 domain organization. Pathogenic AML/MDS variants are indicated. Numbers represent corresponding amino acid positions. **b:** Coomassie-stained gel of 6his-tagged DDX41 full length, R525H, and 153-410 mutant after expression and purification from SF9 insect cells. **c:** Fluorescence polarization (FP) assay of full length DDX41 and indicated 6-FAM-conjugated oligonucleotides in n = 2 independent experiments with individually thawed protein aliquots. The protein concentration on a log2-scale is plotted against the FP in mP (milipolarization unit). Data are represented as mean values +/- standard deviation. Colored lines represent Michaelis-Menten fits. **d:** ADP-Glo assay measuring ATP hydrolysis of full-length DDX41 in n = 5 independent experiments using individually thawed proteins. Titrated DDX41 was incubated together with 100 nM RNA-DNA hybrid substrate with an ssDNA overhang and 5 µM ATP for 60 minutes at 37°C. ATP to ADP conversion was measured by luciferase activity. Results are displayed as the percentage of consumed ATP compared to total ATP, based on an interpolation against a standard curve. Dots indicate individual measurements. Data are represented as mean +/- standard deviation. **e:** FRET-based RNA-DNA hybrid displacement assay. Titrated full-length (FL) DDX41 is incubated with 100 nM

RNA-DNA hybrid substrate and 5 µM ATP. Displacement of the IBFQ-conjugated 38-mer DNA oligo from the 6-FAM-conjugated 13-mer RNA oligo was measured by the change in fluorescence intensity. Data of n = 3 independent experiments with individually thawed proteins are represented as mean values +/- standard deviation (n = 2 for 2.5 µM, unlabeled DNA). **f:** Displacement assay from Figure 13e using titrated full length (FL) DDX41, R525H, or 153-410 mutant. Data are represented as mean +/- standard deviation. Dots indicate results of n = 3 independent experiments with individually thawed proteins. P-values ($p = 0.005$, $p = 0.0021$, $p < 0.0001$, $p = 0.0329$, $p < 0.0001$) derived by two-way ANOVA with Tukey correction for multiple comparisons.

2.4 DDX41 associates with promoters to maintain R-loop homeostasis

Taken together, DDX41 directly bound and unwound RNA-DNA hybrids *in vitro* and prevented global R-loop accumulation *in vivo*, suggesting a potential role as an R-loop resolvase. Unwinding of R-loops *in vivo* would require a direct association with chromatin. To assess whether DDX41 binds specific genomic regions, we performed greenCUT&RUN in a U2OS cell line that expresses GFP-tagged DDX41 under the control of a doxycycline-inducible promoter (Figure 14a). Micrococcal nuclease (MNase) was targeted to DDX41-GFP chromatin binding sites by a GFP nanobody. We confirmed that GFP-tagged DDX41 mirrored the pan-nuclear localization of endogenous DDX41 (Figure 14b). 19,327 greenCUT&RUN peaks were distributed across the genome, of which 6,363 were consistently identified in 2 biologically independent experiments. Ranking of genes bound by DDX41 based on their expression level derived from RNA-sequencing in U2OS cells revealed that DDX41 binding correlated with higher gene expression levels (Figure 14b). Moreover, DDX41 predominantly associated with promoter regions (+/- 3 kb around the TSS), which accounted for 41% of all DDX41 binding sites (Figure 14c, 14d).

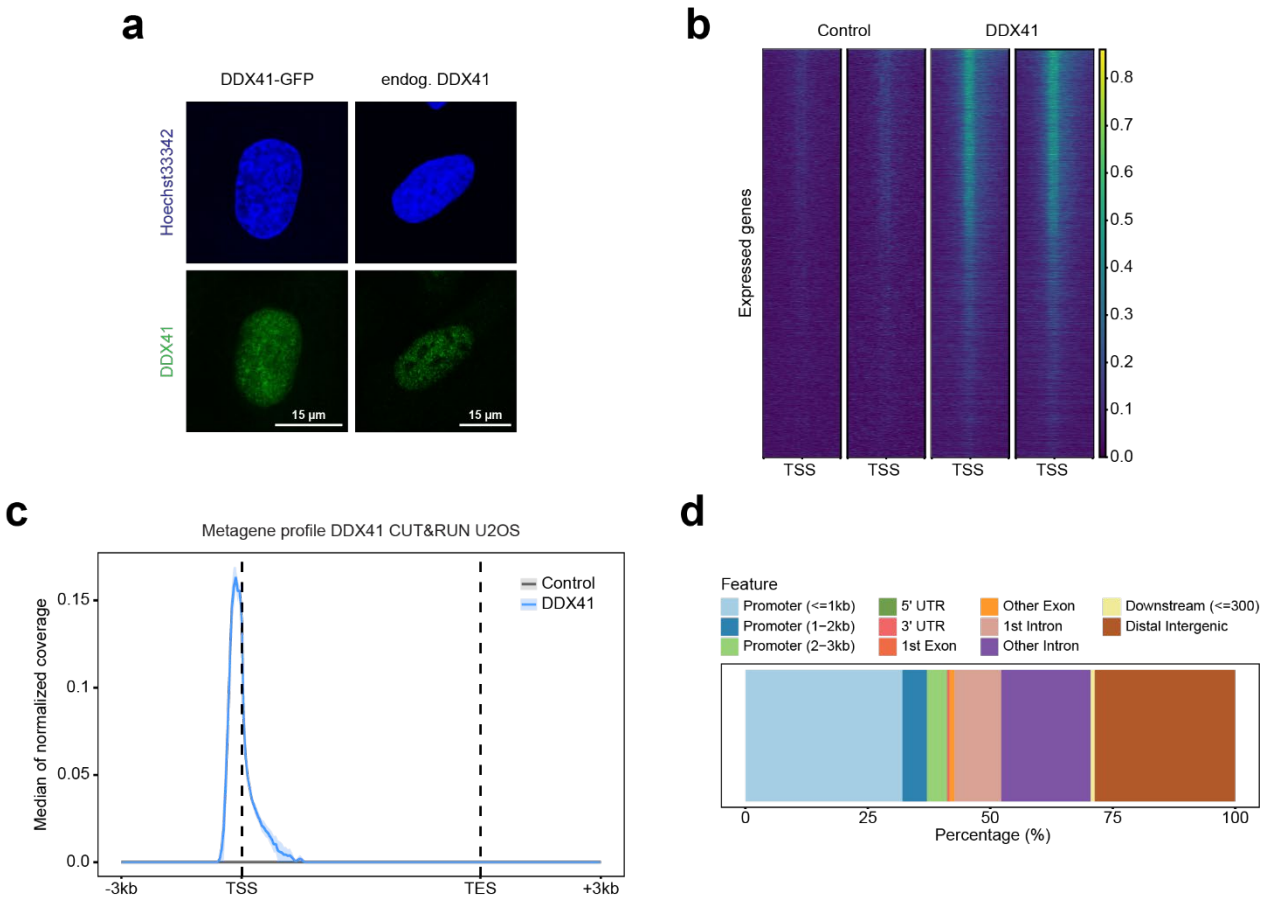


Figure 14: DDX41 binds to the promoter region

a: Representative immunofluorescence images from n = 1 experiment of the subcellular localization of doxycycline-inducible GFP-tagged DDX41 and endogenous (endog.) DDX41. Scale bars represent 15 µm. **b:** Heat maps of greenCUT&RUN signal from -2 kilo bases (kb)/+2kb around the transcription start site (TSS) of expressed genes sorted by expression level. **c:** Metagene profile showing the distribution of the GFP-DDX41 greenCUT&RUN signal in U2OS cells along expressed genes. **d:** Genomic features overlapping GFP-DDX41 greenCUT&RUN peaks in U2OS cells. Features are color-coded as indicated in the legend.

We further investigated if DDX41 chromatin binding sites coincided with regions of R-loop stabilization upon loss of DDX41. To this end, we performed MapR in U2OS cells to identify sites in the human genome with increased R-loop occupancy in the absence of DDX41. MapR is a targeted-nuclease approach that facilitates a catalytically-dead *E. coli* Ribonuclease H to target an MNase to regions on chromatin that are occupied by R-loops under native conditions. R-loop-containing genomic fragments, which are cleaved upon activation of the MNase, are purified and analyzed by next-generation sequencing. MapR analysis revealed a strong correlation between R-loop abundance and gene expression levels based on RNA-sequencing, indicating that R-loop formation was transcription-dependent (Figure 15a). In accordance, inhibition of transcription

using ActinomycinD significantly reduced MapR signal (Figure 15a). We observed global R-loop accumulation in DDX41 knockdown cells, confirming our prior results obtained by S9.6 dot-blot and HBD-GFP retention assay (Figure 15a). Global R-loop accumulation in the absence of DDX41 was also reflected in the number of detected peaks. 35,627 peaks were evident after DDX41 knockdown, while control cells displayed 24,492 R-loop-containing regions. As for DDX41 chromatin binding sites, R-loops in DDX41 knockdown cells were predominantly identified in the promoter region (+/- 3 kb around the TSS) (Figure 15b, 15c). Importantly, promoter-associated R-loops were significantly increased in the absence of DDX41 (Figure 15c, 15d). We quantified 6,810 regions with R-loop gains (FC > 2), while 505 regions displayed R-loop decrease (FC < -2) (Figure 15 e). Moreover, 81% of regions with increased R-loops mapped to the promoter region (+/- 3kb around the TSS) and 74% overlapped with CpG island promoters (CGIs) (Figure 13f, 13g). Of note, none of the regions with R-loop loss overlapped with CGIs, but with intronic and distal intergenic regions (Figure 15f). In total, 5,506 promoters displayed increased R-loop occupancy in DDX41 knockdown cells. Gene set enrichment analysis for Reactome pathways revealed that upon DDX41 depletion R-loops accumulated in genes encoding proteins involved in chromatin organization, NOTCH, and TGF-β signaling (Figure 15h).

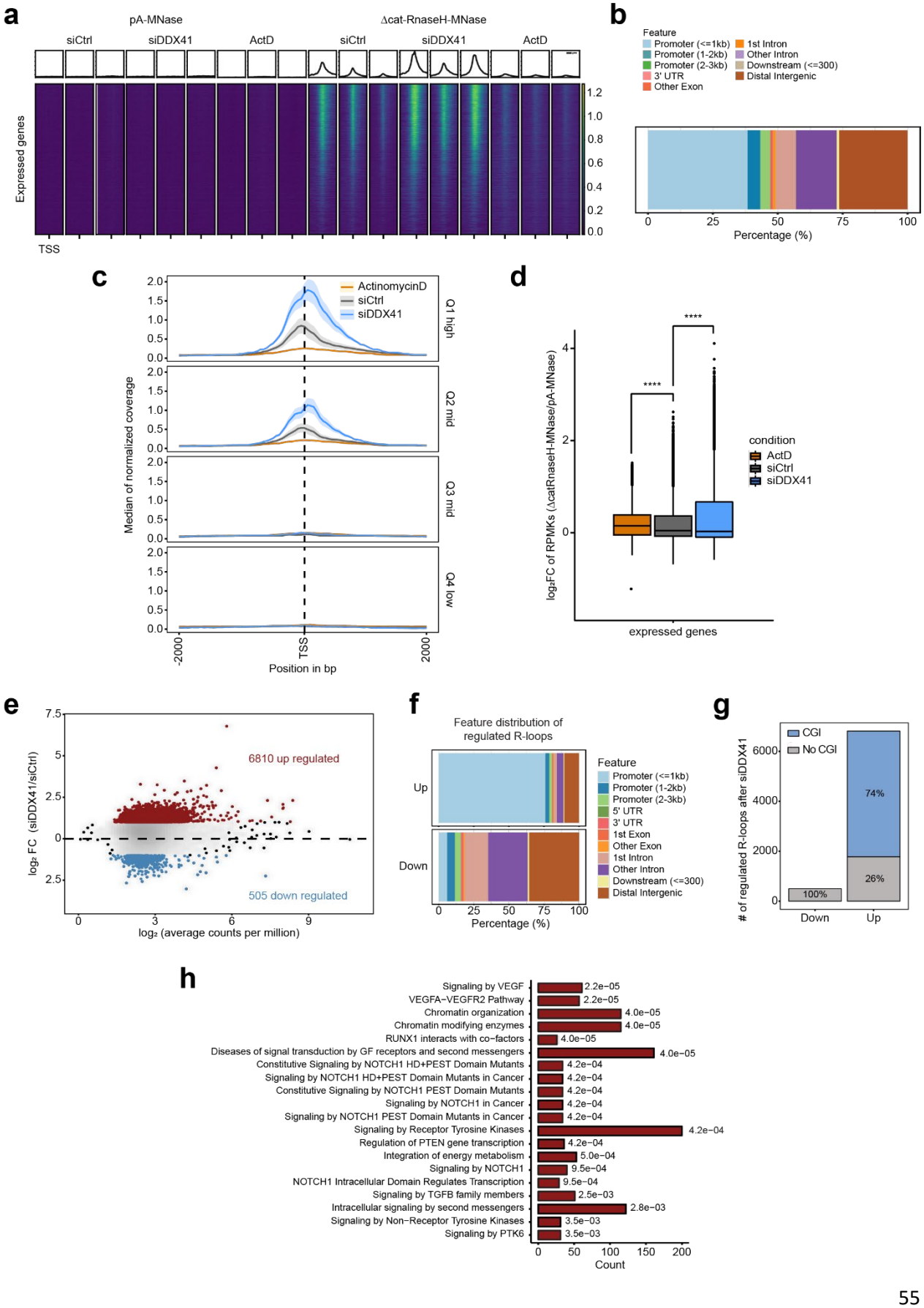

a
pA-MNase
ΔcatRnaseH-MNase
siCtrl
siDDX41
ActD
siCtrl
siDDX41
ActD
Expressed genes
TSS
1.2
1.0
0.8
0.6
0.4
0.2
0.0
b
Feature
Promoter (<=1kb)
Promoter (1-2kb)
Promoter (2-3kb)
3' UTR
Other Exon
1st Intron
Other Intron
Downstream (<=300)
Distal Intergenic
0
25
50
75
100
Percentage (%)
c
ActinomycinD
siCtrl
siDDX41
Q1 high
Q2 mid
Q3 mid
Q4 low
Median of normalized coverage
-2000
TSS
2000
Position in bp
d
****
****
log2FC of RPMKs (ΔcatRnaseH-MNase/pA-MNase)
expressed genes
condition
ActD
siCtrl
siDDX41
e
6810 up regulated
505 down regulated
log2 FC (siDDX41/siCtrl)
log2 (average counts per million)
f
Feature distribution of
regulated R-loops
Up
Down
Feature
Promoter (<=1kb)
Promoter (1-2kb)
Promoter (2-3kb)
5' UTR
3' UTR
1st Exon
Other Exon
1st Intron
Other Intron
Downstream (<=300)
Distal Intergenic
Percentage (%)
g
CGI
No CGI
74%
26%
100%
# of regulated R-loops after siDDX41
Down
Up
h
Signaling by VEGF
VEGFA-VEGFR2 Pathway
Chromatin organization
Chromatin modifying enzymes
RUNX1 interacts with co-factors
Diseases of signal transduction by GF receptors and second messengers
Constitutive Signaling by NOTCH1 HD+PEST Domain Mutants
Signaling by NOTCH1 HD+PEST Domain Mutants in Cancer
Constitutive Signaling by NOTCH1 PEST Domain Mutants
Signaling by NOTCH1 in Cancer
Signaling by NOTCH1 PEST Domain Mutants in Cancer
Signaling by Receptor Tyrosine Kinases
Regulation of PTEN gene transcription
Integration of energy metabolism
Signaling by NOTCH1
NOTCH1 Intracellular Domain Regulates Transcription
Signaling by TGFB family members
Intracellular signaling by second messengers
Signaling by Non-Receptor Tyrosine Kinases
Signaling by PTK6
2.2e-05
2.2e-05
4.0e-05
4.0e-05
4.0e-05
4.0e-05
4.2e-04
4.2e-04
4.2e-04
4.2e-04
4.2e-04
4.2e-04
4.2e-04
5.0e-04
9.5e-04
9.5e-04
2.5e-03
2.8e-03
3.5e-03
3.5e-03
Count

Figure 15: DDX41 opposes excessive R-loop formation in the promoter region of highly transcribed genes

a: MapR performed in three biological replicates in U2OS cells after 48 h knockdown with control siRNA, DDX41 siRNA, or treatment with 4 µM ActinomycinD D for 6 hours. Heat maps of normalized read coverage ranging from -/+2kb around the transcription start site of expressed genes sorted by gene expression based on the RNA-sequencing analysis of U2OS cells. **b:** Bar plot displaying the feature distribution of all peaks called by MACS2 in the MapR experiment in n = 3 biologically independent experiments from U2OS cells after 48 h knockdown with control siRNA. Consensus peaks were constructed using the intersection of peaks in the replicates per group (control, DDX41 knockdown) and the union of the resulting peaks was used as a consensus peak set. The proportion of the individual feature is represented in the indicated colors. **c:** MapR in U2OS cells performed in n = 3 biologically independent experiments after 48 h knockdown with control siRNA (grey), DDX41 siRNA (blue), or treatment with 4 µM actinomycinD D for 6 hours (orange). Metagene profiles of MapR signal ranging from -2kilo bases (kb)/+2kb around the transcription start site (TSS) of all expressed genes based on RNA-sequencing analysis of U2OS cells, divided into 4 quantiles. Shadows indicate the standard error of the median between replicates. **d:** The $\log_2$ fold change (FC) of the reads per kilo base per million mapped reads (RPMK) of MapR signal between the ΔcatRNaseH-MNase and the untargeted pA-MNase obtained +/-1kb around the transcription start site (TSS) of all expressed genes based on RNA-sequencing. Mean RPMKs between n = 3 biologically independent experiments are shown. Whiskers represent the 5th-95th percentile, limits the 25th-75th percentile, the center line the median, and individual dots depict outliers. P-values ($p < 0.0001$, $p < 0.0001$) were derived using two-sided paired Welch´s t-test. **e:** Scatter plot of MapR regions in U2OS cells. Consensus regions were constructed using the intersection of peaks for the replicates in each condition (siCtrl and siDDX41). The union of these regions was used for further analysis and quantification of the coverage/FC. The mean $\log_2$ fold change between siCtrl and siDDX41 is plotted against the $\log_2$ average counts per million representing the coverage. Genomic regions that are differentially regulated (FC>2) are highlighted in red (up) or in blue (down). **f:** Genomic feature distribution of the regulated MapR regions in U2OS cells after DDX41 knockdown. Features are color-coded as indicated in the legend. **g:** The proportion of genomic regions with R-loop gains or losses in U2OS cells overlapping CGIs or not-overlapping regions are depicted. **h:** Reactome pathway over-representation analysis for genes with R-loop gains in U2OS cells. The adjusted p-values (Fisher´s exact test with Bonferroni-Holm correction) are indicated.

We overlapped promoters bound by DDX41 with promoters that displayed R-loop gains in DDX41 knockdown cells. Binding of DDX41 coincided with increased R-loop formation at the exemplary *USP1* promoter (Figure 16a). Globally, roughly 39% of promoters with R-loop increase were also occupied by DDX41 (Figure 16b). This overlap might be an underestimation since DDX41 is loosely associated with chromatin, and greenCUT&RUN was performed with GFP-tagged DDX41 while the endogenous DDX41 was still present. Taken together, our data strongly suggest a potential role for DDX41 in maintaining promoter R-loop homeostasis. Promoter R-loops, especially within CGIs, have been associated with transcriptional regulation[73]. We monitored global nascent transcriptional activity by measuring 5-Ethynyl Uridine (EU) incorporation. Intriguingly, the accumulation of R-loops in the promoter region in DDX41 knockdown cells did not cause a global transcription defect (Figure 16c). However, we found a significantly decreased serine 5- and significantly increased serine 2-phosphorylation of the RBP1 C-terminal domain (CTD), suggesting altered transcriptional dynamics (Figure 16d, 16e), potentially, due to

augmented RNA Pol II promoter-proximal pausing due to excessive R-loop formation around the transcription start site.

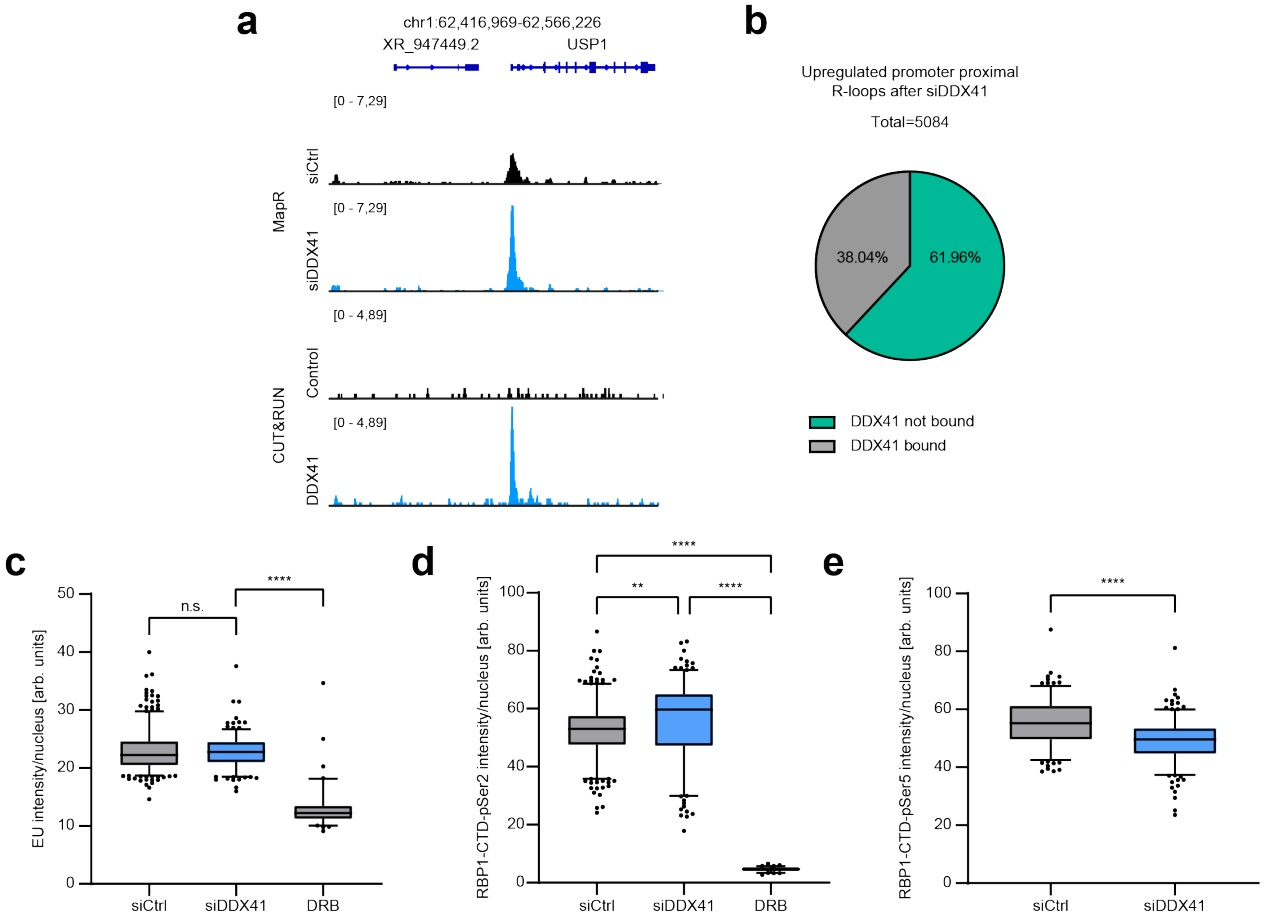


Figure 16: Promoter R-loop gains in DDX41 knockdown cells overlap with DDX41 binding sites

a: Representative snapshot of a genomic region depicting R-loops and GFP-DDX41 binding profiled by MapR and greenCUT&RUN, respectively, in U2OS cells. **b:** Pie chart shows the proportion of R-loop gains in promoter regions that were also found to associate with GFP-DDX41 in CUT&RUN experiments in U2OS cells. **c:** Confocal microscopy analysis of EU intensity in U2OS cells after transfection for 48 h with control or DDX41 siRNA or DRB treatment for 3 h. Center-line within the box represents the median, whiskers the 5th-95th percentile, limits the 25th-75th percentile and dots indicate outliers. At least 200 cells per condition were measured across n = 1 experiment. P-values ($p = 0.761$, $p < 0.0001$) derived using one-Way-ANOVA with Tukey correction for multiple comparisons. **d/e:** Confocal microscopy analysis of RBP1-CTD pSer2 or pSer5 intensity in U2OS cells after transfection for 48 h with control or DDX41 siRNA or DRB treatment for 3 h. The Center line within the box represents the median, whiskers the 5th-95th percentile, limits the 25th-75th percentile and dots indicate outliers. At least 200 cells per condition were measured across n = 1 experiment. P-values ($p = 0.0095$, $p < 0.0001$, $p < 0.0001$) were calculated using one-way-ANOVA with Tukey correction for multiple comparisons. Ser5: p-value < 0.0001 derived using two-sided student´s t-test.

2.5 DDX41 opposes promoter DSB hotpots and inflammatory signaling

Global R-loop accumulation in DDX41 knockdown cells led to R-loop-dependent genomic instability. We further investigated whether the genomic instability is a result of DSB formation and if DSBs co-localized with R-loop hotspots in the genome. To map DSBs genome-wide after DDX41 depletion, we performed Break Labeling In Situ and Sequencing (sBLISS) in HCT116 cells[354,355]. As previously described[354,356–358], spontaneous DNA fragility hotspots predominantly overlapped with promoter regions in unperturbed cells (Figure 17a, 17b, 17c). The tendency towards promoter fragility was even more striking in DDX41 knockdown cells, in which 63% of hotspots overlapped with promoter regions (Figure 17a, 17b, 17c). In total, we identified 8,381 DNA fragility hotspots in the absence of DDX41. Of those, 2,423 were also present in unperturbed cells, whereas 5,958 were exclusively identified in DDX41 knockdown cells (Figure 17d). To assess to which extent R-loop accumulation coincided with DNA fragility hotspots, we performed MapR in HCT116 cells. Similar to U2OS cells, R-loop levels were dependent on the transcriptional status of a gene, which we determined based on published GRO-sequencing data from HCT116 cells (Figure 17e)[359]. Loss of DDX41 in HCT116 resulted in significantly increased R-loops in the promoter region of transcribed genes (Figure 17e). Within 3 biologically independent experiments, we identified 15,177 R-loop-containing regions in DDX41 knockdown HCT116 cells, of which 7,275 displayed a gain in R-loops (FC > 1.5) compared to unperturbed control cells (Figure 17f). R-loop stabilization and increased DNA fragility coincided at the exemplary *NDFIP2* promoter in DDX41 knockdown cells (Figure 17g). Across the genome, 1,642 promoter regions displayed increased DNA fragility in the absence of DDX41 based on sBLISS. Of those, 53% also showed excessive R-loop formation after DDX41 depletion (Figure 17h).

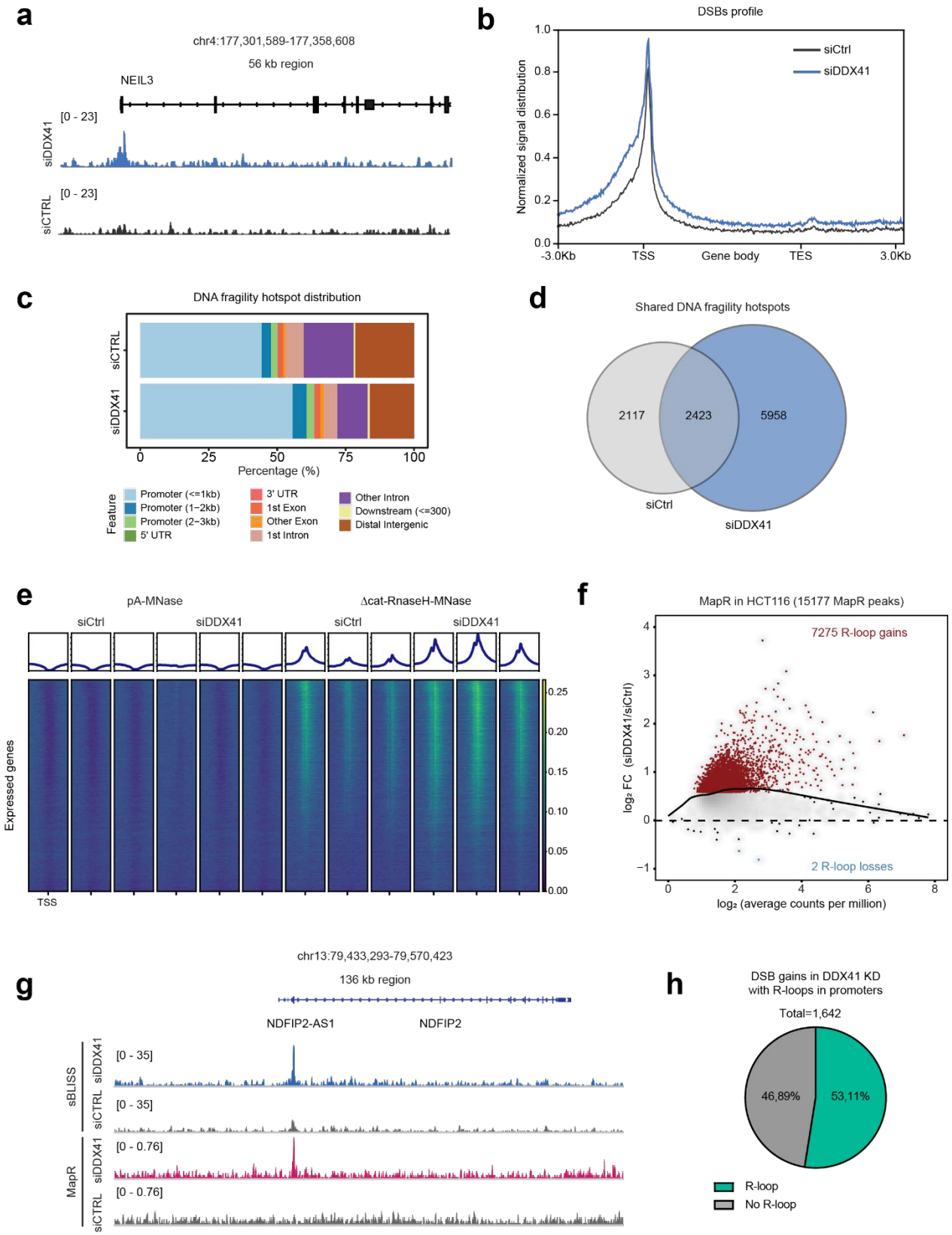

a
chr4:177,301,589-177,358,608
56 kb region
NEIL3
[0 - 23]
siDDX41
[0 - 23]
siCTRL
b
DSBs profile
Normalized signal distribution
siCtrl
siDDX41
-3.0Kb
TSS
Gene body
TES
3.0Kb
c
DNA fragility hotspot distribution
siCTRL
siDDX41
Percentage (%)
Feature
Promoter (<=1kb)
Promoter (1–2kb)
Promoter (2–3kb)
5' UTR
3' UTR
1st Exon
Other Exon
1st Intron
Other Intron
Downstream (<=300)
Distal Intergenic
d
Shared DNA fragility hotspots
2117
2423
5958
siCtrl
siDDX41
e
pA-MNase
ΔcatRnaseH-MNase
siCtrl
siDDX41
siCtrl
siDDX41
Expressed genes
TSS
f
MapR in HCT116 (15177 MapR peaks)
7275 R-loop gains
2 R-loop losses
log₂ FC (siDDX41/siCtrl)
log₂ (average counts per million)
g
chr13:79,433,293-79,570,423
136 kb region
NDFIP2-AS1
NDFIP2
sBLISS
siDDX41
[0 - 35]
siCTRL
[0 - 35]
MapR
siDDX41
[0 - 0.76]
siCTRL
[0 - 0.76]
h
DSB gains in DDX41 KD
with R-loops in promoters
Total=1,642
46,89%
53,11%
R-loop
No R-loop

Figure 17: DDX41 opposes deleterious R-loop accumulation and associated DNA fragility in the promoter regions

a: Representative snapshot of the *NEIL3* genomic region depicting DNA fragility profiled by sBLISS in wild type and DDX41 knockdown HCT116 cells. **b:** Metagene profile showing the double-strand break (DSB) signal distribution profiled by sBLISS along genes in wild type and DDX41 knockdown HCT116 cells. **c:** Genomic feature overlapping DNA fragility hotspots mapped by sBLISS in wild type and DDX41 knockdown HCT116 cells. Features are color-coded as indicated in the legend. **d:** Venn diagram showing the number of unique and overlapped peaks mapped by sBLISS in wild type and DDX41 knockdown HCT116 cells. **e:** Heat maps showing the distribution of the MapR signal of all expressed genes for pA-MNase and ΔcatRNaseH-MNase conditions after 48 h control or DDX41 knockdown in HCT116 cells. MapR signal was sorted based on gene expression level obtained from the previously published GRO-sequencing analysis from high to low. The region -2kb/+ 2kb around the TSS is individually displayed for each biological replicate. Metagene profiles are outlined above the associated heat maps. **f:** Scatter plot of MapR regions in HCT116 cells. Consensus regions were constructed using the intersection of peaks for the replicates in each condition (siCtrl and siDDX41). The union of these regions was used for further analysis and quantification of the coverage/fold change (FC). The mean $\log_2$ FC between siCtrl and siDDX41 is plotted against the $\log_2$ average counts per million representing the coverage. Genomic regions that are differentially regulated (FC>2) are highlighted in red (up) or in blue (down). **g:** Representative snapshot of a genomic region showing accumulation of R-loops and DSBs profiled by MapR and sBLISS, respectively, in HCT116 cells. h: Pie chart showing the percentage of double-strand breaks (DSB) gains (FC >2) mapped to promoters in DDX41 knockdown HCT116 cells that overlap or not with R-loops mapped in DDX41 knockdown (KD) HCT116 cells.

More than half of the increased DNA fragility sites coincided with the increased R-loop formation in DDX41 knockdown cells, suggesting that DDX41 is required to prevent R-loop-dependent genomic instability in the promoter region. Excessive R-loop formation due to dysregulation and subsequent DNA damage have been reported as potent activators of an inflammatory response. Strikingly, RNA-sequencing after knockdown of DDX41 in U2OS cells revealed upregulation of inflammatory signaling genes, in particular NF-kB signaling (Figure 18a). Accordingly, we observed significantly increased nuclear localization of the NF-kB subunit p65 upon loss of DDX41, confirming an active inflammatory response (Figure 18b). Of note, NF-kB genes were neither occupied by DDX41, nor displayed changes in R-loop levels after DDX41 knockdown. Activation of inflammatory signaling was most likely triggered by increased genomic instability and not due to changes in transcription via altered R-loop homeostasis in the promoter region of inflammatory response genes.

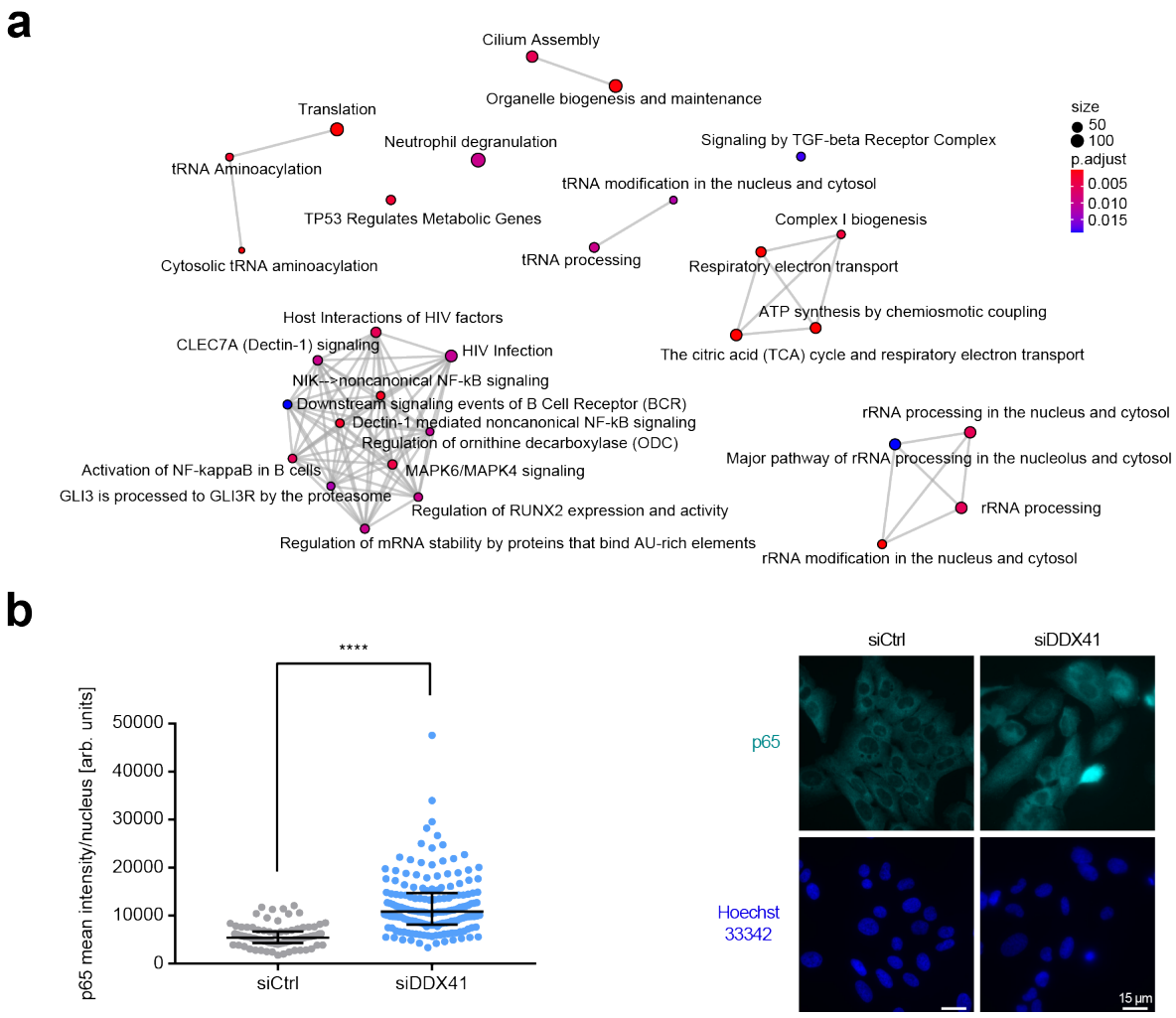


Figure 18: R-loop-dependent genomic instability in DDX41 knockdown cells is accompanied by an inflammatory response

a: Network of the Reactome pathway enrichment analysis of up-regulated genes after DDX41 knockdown compared to control knockdown based on RNA-seq. All expressed genes were used as background. The size of the dots indicates the number of genes contributing to the displayed term. Gradual coloring represents the adjusted p-values based on Fisher´s exact test with Bonferroni correction for multiple comparisons. **b:** (left) Immunofluorescence analysis of p65 after 48 h of indicated knockdowns in U2OS cells. Dots represent measurements of individual cells, black line indicates the median of the population with interquartile range. Representative data of n = 2 biologically independent experiments. P-value < 0.0001 derived from n > 100 cells using unpaired, two-sided Student's t-test. (Right) Confocal microscopy of p65 after 48 h control knockdown or DDX41 knockdown in U2OS cells. Representative images of p65 staining (cyan) and DNA counterstaining with DAPI are displayed on the right. Quantification is shown on the left. Data derived from n = 2 biologically independent experiments. Scale bars represent 15 µm.

2.6 DDX41 opposes spontaneous DSB formation in HSPCs

Pathogenic germline and somatic mutations of DDX41 have been identified in MDS and AML patients[264,270,360]. To assess the potential role of R-loop-dependent genomic instability in the development of MDS and AML, we depleted DDX41 in hematopoietic stem and progenitor cells

(HSPCs) and monitored 53BP1 foci as a proxy for spontaneous DSB formation. DDX41 knockdown HSPCs displayed significantly increased 53BP1 foci, likely due to excessive R-loop formation (Figure 19a). The expression of pathogenic DDX41 AML variants (R525H and L237F+238T) in HSPCs led to a mild increase in 53BP1 foci formation compared to the expression of wild type DDX41. These findings suggest that wild-type DDX41 is required to maintain genomic stability and that AML variants of DDX41 act as dominant-negative mutants (Figure 19b).

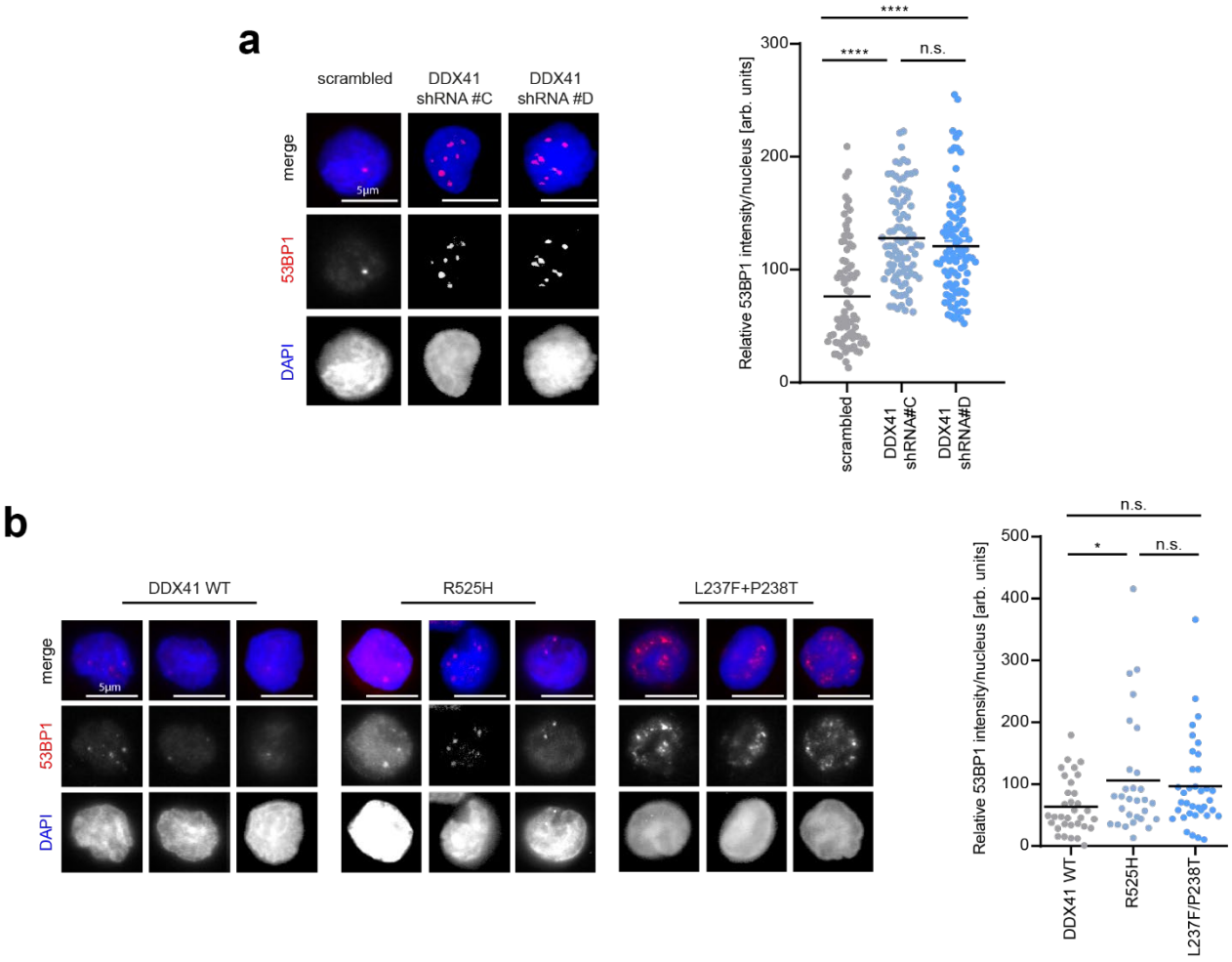


Figure 19: Loss of DDX41 or expression of pathogenic AML variants induces DSB formation in HSPCs

a: Immunofluorescence analysis of HSPCs after 24 h of indicated knockdowns. Cells were nucleofected with plasmids encoding GFP and respective shRNAs. GFP-positive cells were sorted via FACS and seeded on coverslips. Representative images of 53BP1 (red) staining in HSPCs (left). DNA was counterstained with DAPI (blue). Quantification of nuclear 53BP1 intensity (right). Each dot represents a single measured value. The black line indicates the median. At least 80 cells across n = 2 biologically independent experiments were measured per condition. P-values ($p < 0.0001$, $p < 0.0001$, $p = 0.2227$) were calculated by one-way ANOVA with Tukey correction for multiple comparisons. Scale bars - 20 µm. **b:** Immunofluorescence analysis of HSPCs after expression of DDX41 WT, L237F+P238T, or R525H mutants tagged with GFP. GFP-positive cells were sorted by FACS and used for the analysis. Representative images of 53BP1 (red) staining in HSPCs (left). DNA was counterstained with DAPI (blue).

Quantification of nuclear 53BP1 intensity (right). Dots represent results from individual cells. The median is indicated by the black line. P-values ($p = 0.0237$, $p = 0.624$, $p = 6.018$) were derived from at least 30 cells across $n = 1$ experiment using one-way ANOVA with Tukey correction for multiple comparisons. Scale bars - 20 µm.

We propose that wild-type DDX41 associates with promoter regions of highly transcribed genes in human cells to maintain promoter-proximal R-loop homeostasis. Unwinding of promoter-proximal R-loops promotes faithful RNA Pol II transcription elongation. In the absence of functional DDX41 or the presence of pathogenic MDS/AML variants, impaired promoter-proximal R-loop unwinding leads to R-loop accumulation, replication stress, and promoter-associated DSB hotspots. Cells lacking DDX41 experience R-loop-dependent DNA damage and replication stress, acquire dependency on ATR signaling, and activate an inflammatory response (Figure 20).

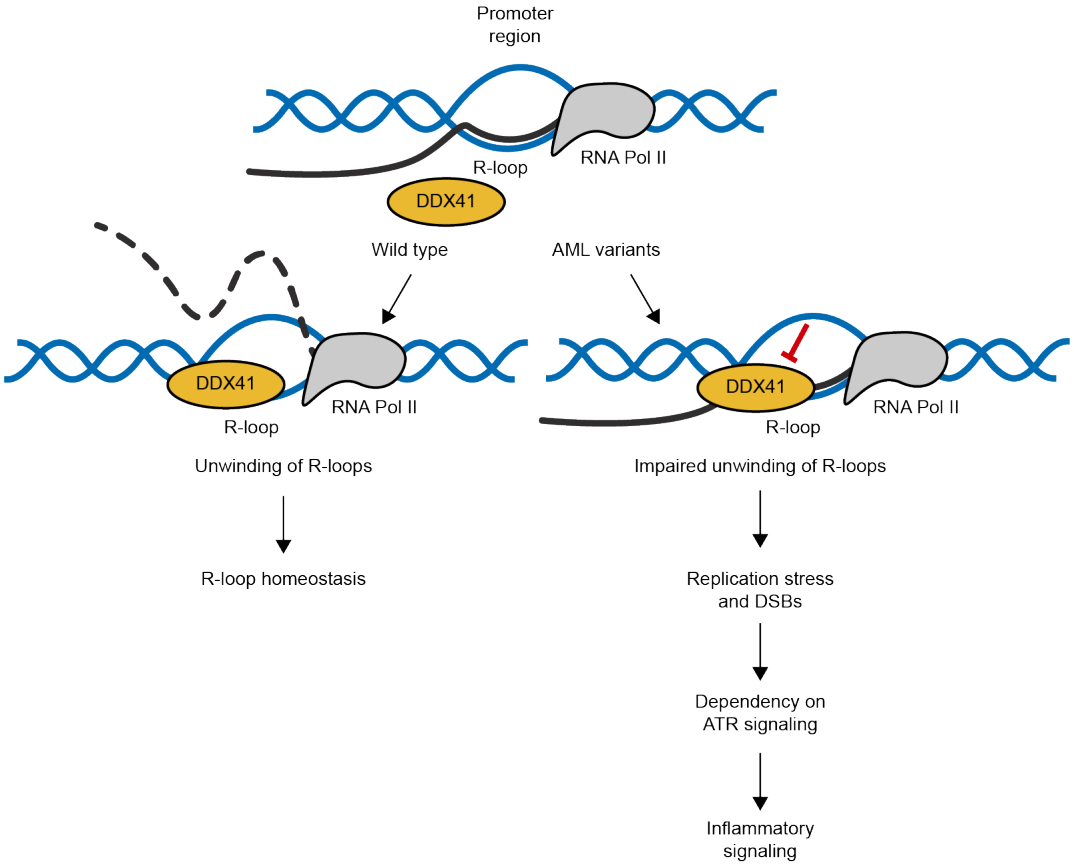


Figure 20: Impaired unwinding of promoter-associated R-loops in the presence of DDX41 AML variants results in replication stress, DSB formation, dependency on ATR, and inflammatory signaling

Wild type DDX41 associates with R-loops in promoters of active genes and balances R-loop levels by unwinding RNA-DNA hybrids. Pathogenic DDX41 variants found in acute myeloid leukemia (AML) display impaired RNA-DNA hybrid unwinding activity, leading to the accumulation of R-loops at promoters. Accumulation of R-loops at promoters results in increased replication stress, DSBs, and inflammatory signaling, rendering DDX41 mutated AML cells vulnerable to ATR inhibitors.

3 Discussion

3.1 Strengths and weaknesses of the RDProx workflow

R-loops contribute to the regulation of various crucial processes such as transcription, mitosis, and DNA repair. Tight control of R-loop abundance is important for cell survival since deleterious R-loop accumulation is associated with replication stress and genomic instability. Mutations in R-loop regulating proteins predispose patients to neurological disorders and cancer[210,250,361]. Understanding the contribution of R-loop dysregulation to disease development requires more precise characterization of the R-loop regulatory proteome. We established an MS-based workflow called RDProx to identify R-loop proximal proteins in human cells. RDProx relies on the hybrid binding domain of RNaseH1, which targets the conjugated ascorbate peroxidase APEX2 to RNA-DNA hybrids on chromatin under native conditions. Specificity towards RNA-DNA hybrids is achieved by utilizing a non-binding mutant of the HBD as a control, which contains three point mutations (HBD-WKK). We applied RDProx in HEK293T cells in a SILAC-based approach and identified 612 R-loop proximal proteins, which we distinguished into two categories depending on their enrichment (Tier 1 and Tier 2). Thus, RDProx provides an important resource for further functional characterization of R-loop regulating proteins.

Before this study, two distinct proteomic approaches have been established to identify R-loop binding proteins. The first study employed the S9.6 antibody to enrich R-loops and associated proteins after cell lysis and chromatin fragmentation, while the second probed biotinylated RNA-DNA hybrid oligo substrates with nuclear extracts to identify R-loop interacting proteins *in vitro*[152,214]. RDProx is in many regards advantageous over the previous two approaches: (1) R-loop proximal proteins are labeled *in vivo* under native conditions before cell lysis, while chromatin and other cellular compartments are still intact; (2) harsh extraction conditions to solubilize chromatin-associated proteins, which are detrimental to R-loops and not applicable to Co-IP-based methods, can be used after biotinylation; (3) high salt washing conditions are applicable due to the strong binding between biotin and NeutrAvidin during the enrichment step, thus depleting unspecific secondary interactors during the pulldown; (4) high discrimination between RNA-DNA hybrids and RNA species by using an S9.6-independent probe to target APEX2 to R-loops[362]. Unspecific binding of the S9.6 antibody to RNA species, especially rRNA precursors, has recently raised doubts in S9.6-based approaches[362]. Indeed, an unbiased comparison between RDProx and the

S9.6-based proteomic study by gene set enrichment analysis revealed “rRNA processing” and “ribosome biogenesis” as the most significant terms for proteins, exclusively enriched by S9.6 co-IP (Figure 8a). Similar terms were significantly enriched for proteins identified by the *in vitro* RNA-DNA hybrid pulldown, which is inherently biased towards highly abundant proteins (Figure 8b). Terms associated with ribosome biogenesis were not significantly enriched among the proteins identified by RDProx since the HBD discriminates better between RNA-DNA hybrids and RNA than the S9.6 antibody. Furthermore, since RDProx is based on proximity labeling, it is less biased towards highly abundant proteins such as ribosomal proteins compared to *in vitro* pulldowns. However, also RDProx possesses inherent drawbacks. The HBD-APEX2 fusion protein needs to be exogenously expressed, which makes RDProx inapplicable for the identification of R-loop proximal proteins in primary cells and patient samples. Moreover, RDProx relies on the HBD of RNaseH1 and is therefore inherently biased to the subset of R-loops that is bound by RNaseH1. Comparison of S9.6 or RNaseH1-based genome-wide R-loop mapping approaches suggests the existence of two classes of R-loops[72]. S9.6-based methods recover mostly long R-loops within the gene body and the transcription termination site, while RNaseH1-based workflows mostly identify short promoter-proximal R-loops, which frequently form during promoter-proximal pausing of RNA Pol II. In contrast, an alternative study suggests that the bias arises due to the distinct protocols of *ex vivo* and *in vivo* approaches and not due to different binding properties of the probes[256]. Promoter-proximal R-loops were identified under native conditions, while *ex vivo* mappings predominantly led to the recovery of R-loops in the gene body and the transcription termination site[256]. Intriguingly, this was independent of the probe, suggesting that RNaseH1 binding is not restricted to promoter-proximal R-loops. As of now, it remains unclear whether the HBD of RNaseH1 only binds to a specific subset of R-loops in the genome. However, since RDProx is performed *in vivo* under native conditions, we acknowledge that RDProx might recover mostly proteins that associate with promoter-proximal R-loops. Accordingly, RDProx identified proteins involved in splicing regulation and spliceosome assembly, which could present from loading of the spliceosome onto the nascent transcript during promoter-proximal pausing of RNA Pol II (Figure 8a). Accordingly, we also found multiple components of the mediator complex and transcription factors which are associated with promoter regions (Figure 8a). Moreover, we detected a number of proteins involved in DNA replication that might indicate conflicts of the transcription and the replication machinery or overlap of transcription initiation sites and replication origins at CGIs.

3.2 RDProx deciphers the human R-loop proximal protein network

We successfully mapped the human R-loop proximal proteome by applying our newly established RDProx workflow in human cells. The vast majority of R-loops in the human genome form co-transcriptionally in association with RNA polymerases[255,363]. Transcription-dependent R-loop formation is not exclusive to protein-coding genes transcribed by RNA Pol II but has also been observed for nucleolar rRNA loci and tRNA loci, transcribed by RNA Pol I and RNA Pol III, respectively[93]. In accordance, we identified several subunits from the three different RNA polymerase complexes such as POLR1A, POLR2A/B, and POLR3A/F, underlining the co-transcriptional character of the structure. Importantly, RDProx does not appear to be biased to RNA Pol II-transcribed protein-coding genes but also maps proximal proteins of RNA Pol I and RNA Pol III-associated R-loops. Moreover, we found multiple subunits of the mediator complex (MED1, MED12, MED13) and various transcription factors (GTF2E1, GTF2I, GTF3C2, BCLAF1, NFYC), which together with the identification of spliceosome assembly proteins suggests that RDProx preferentially identifies proteins proximal to promoter-associated R-loops (Figure 8a).

RNA modifications define an important feature for the regulation of gene expression. One of the most frequent covalent RNA modifications is the N^6-methyladenosine (m^6A)[364]. Multiple recent studies have reported m^6A-modified RNA within RNA-DNA hybrids[239–241]. Our RDProx analysis provides first evidence for the association of m^6A RNA machinery with R-loops in human cells. Alongside m^6A writers (m^6A-METTL-associated complex: VIRMA, ZC3H13, and RBM15), we identified readers (HNRNPA2B1 and HNRNPC) and erasers (ALKBH5 and FTO) near R-loops (Figure 8a). These identifications suggest that m^6A-modified RNA within R-loops displays an additional layer for the regulation of R-loop stability through RNA processing proteins. Indeed, m^6A deposition on RNA-DNA hybrids has been shown to influence R-loop homeostasis. During mitosis, m^6A deposition on RNA-DNA hybrids by METTL3 leads to the recruitment of the m^6A reader YTHDF2, which subsequently promotes R-loop turnover[240]. While mitotic R-loops seem to be destabilized in an m^6A-dependent manner, m^6A deposition within R-loops at transcription termination sites increases R-loop stability[239]. Loss of m^6A within RNA-DNA hybrids and causal R-loop destabilization at transcription termination sites causes RNA Pol II read-through transcription due to aberrant transcription termination[239]. Further studies are required to unravel the role of dynamic m^6A deposition within different classes of R-loops.

DNA replication-associated proteins define another large group, which is proximal to R-loops based on RDProx. We identified components of the MCM complex, WDHD1, RFC1/2/4, MSH6, CHAF1B, RECQL, and the ssDNA-binding proteins RPA1/2. Conflicts between the replication machinery and the transcription machinery in the form of R-loops have been shown to negatively influence replication fork progression and promote DNA damage. Accordingly, RNase H enzymes ensure unperturbed replication fork progression by removing unscheduled R-loops[230]. Different orientations of the replication machinery and transcription machinery during TRCs have been shown to influence R-loop formation due to changes in DNA supercoiling. Co-directional conflicts lead to a reduction in R-loops, whereas head-on conflicts result in R-loop stabilization[172]. Moreover, the MCM complex itself may act as an R-loop resolvase, since it possesses *in vitro* helicase activity towards RNA-DNA hybrids[365]. Therefore, the MCM helicase complex could potentially remove co-directional R-loops before the replication fork during S-phase to prevent TRCs. ATAD5-dependent PCNA unloading behind the replication fork and recruitment of DEAD-box helicases to replication forks have also been described as two mechanisms to prevent collisions between the replisome and the transcription machinery[366]. Moreover, co-occurrence of replication origins and transcription start sites at CGIs requires tight regulation and timely coordination of the two processes to prevent TRCs[188,367]. Since replication and transcription share the same template during S-phase, crosstalk between the two dedicated machineries is required to maintain genome stability. RDProx provides further evidence of replication-associated proteins in the proximity of R-loops, thus establishing the foundation for further studies to investigate which proteins oppose TRC-dependent genomic instability in human cells.

Another set of proteins identified by RDProx are chromatin remodelers such as SMARCA4, ARID1A, SMARCC1, and SMARCE1. Their association suggests an R-loop regulatory role for SWI/SFN complexes, which are known to associate with the H3K27 acetylation mark at actively transcribed sites[368,369]. In accordance, chromatin remodeling has been proposed as a central mechanism for R-loop regulation. Loss of ATPase activity of the SWI/SFN complex results in R-loop accumulation, higher TRC frequency, and R-loop-dependent genomic instability, indicating that the SWI/SFN chromatin remodeling activity is crucial to maintaining R-loop homeostasis[370]. Moreover, ARID1A-containing BAF complexes have been shown to prevent replication stress by opposing excessive R-loop accumulation in a TOP2A-dependent manner[371]. In addition to R-loop regulatory functions of chromatin remodelers, R-loop formation can also impact the recruitment of

chromatin remodelers. R-loop stabilization at 5´ends of transcribed genes is associated with the recruitment of the Tip60-p400 chromatin-activating complex, whereas recruitment of the Polycomb repressive complex (PRC2) is inhibited in an R-loop-dependent manner[372]. The identification of components of SWI/SFN complexes by RDProx further implies a potential role for chromatin remodelers at R-loops. Moreover, the presence of the BPTF subunit at R-loops suggests recruitment of the NURF-1 and NURF-5 ISWI chromatin remodeling complexes to R-loops[373].

A significantly overrepresented class of proteins identified at R-loops by RDProx are DEAD/DEAH-box helicases such as DDX5, DDX17, DDX18, DDX19, DDX21, and DDX39B (Figure 8b). One of the most studied R-loop resolvases is the DDX5 helicase[183,216,353,374,375]. DDX5 resolves R-loops at transcription termination sites in a PRMT5-dependent manner through association with Thrap3 and XRN2[216,374]. Moreover, DDX5 is recruited to DSBs by BRCA2 to remove RNA-DNA hybrids at the break site, thereby allowing resection and DNA repair via the HR pathway[183,353]. In contrast, a recent report suggests that DDX17 promotes the formation of RNA-DNA hybrids at DSBs to promote DNA repair, indicating a dynamic role of RNA-DNA hybrids during DSB repair[376]. Very little is known about the role of DDX18 in R-loop regulation. Until today, DDX18 has only been shown to bind RNA-DNA hybrids *in vitro* and to co-localize with R-loops in the nucleoli[214]. Since the co-localization has been identified using the S9.6 antibody, the co-localization may be an artifact and instead be a co-localization with rRNA precursors[362]. Identification of DDX18 by RDProx displays further evidence of its association with R-loops. Another DEAD-box helicase with RNA-DNA hybrid unwinding activity is the DDX21 helicase. De-acetylation by SIRT7 augments the helicase activity of DDX21 in vivo, thereby promoting R-loop unwinding[215]. Through its RNA-DNA unwinding activity, DDX21 ensures RNA polymerase elongation by resolving R-loop-mediated stalling. Loss of DDX21 is associated with increased R-loop-dependent genomic instability. However, which genomic regions are regulated by DDX21 remains elusive. The nucleopore-associated helicase DDX19 re-localizes to chromatin in response to replication stress in an ATR-dependent manner to resolve R-loops at sites of TRCs[217]. Accordingly, DDX19 was not identified by RDProx since it was performed under unperturbed conditions. We also identified the DDX39B/UAP56 helicase in the vicinity of R-loops, which associates with elongating transcription complexes as far as the transcription termination site. Faithful transcription elongation and termination are ensured by R-loop unwinding activity of

DDX39B throughout the gene body until the nascent transcript is released[219]. However, excessive R-loop formation in the absence of DDX39B might also result from alterations in nuclear mRNA export since DDX39B is part of the TREX complex[377,378]. Further studies are essential to differentiate a potential role as an R-loop resolvase from its role in mRNA export. RDProx further supports the idea that DDX39B associates with R-loops in human cells. In addition to previously reported R-loop-associated DEAD-box helicases, our study revealed 5 additional DEAD-box helicases without a known role in R-loop regulation (DDX27, DHX37, DDX39A, DDX41, and DDX42) to be associated with native R-loops.

3.3 DDX41 prevents spontaneous R-loop-dependent DNA damage

Enrichment of DEAD/DEAH-box proteins DDX27, DHX37, DDX39A, DDX41, and DDX42 at R-loops based on RDProx prompted us to investigate their potential role in R-loop metabolism. None of the five helicases has yet been linked to R-loops regulation. Instead, DDX27 has been reported to be part of the rRNA processing machinery[379], DHX37 to participate in ribosome biogenesis[380], DDX39A to be involved in nuclear mRNA export[381,382], DDX41 to be associated with the spliceosome[262] and DDX42 to act as an RNA chaperone[383]. Global R-loop accumulation has been frequently shown to result in spontaneous DNA damage, thus we investigated if knockdown of any of the five helicases led to increased yH2AX levels. We only observed significant yH2AX formation after loss of DDX41, whereas no spontaneous DNA damage was evident after knockdown of the other four helicases (Figure 9a, 9b). Although we cannot rule out that the four other helicases regulate R-loops locally, our data suggest that exclusively DDX41 opposes global R-loop accumulation, thereby safeguarding genomic stability. Overlap with a genome-wide siRNA screen supports our finding, since out of the five helicases only DDX41 ranks in the most significant category for inducing spontaneous yH2AX upon knockdown[384]. Furthermore, DDX41 knockdown led to significantly increased 53BP1 foci under pre-extraction conditions, indicating DSB formation (Figure 9c). R-loop-dependent DNA damage has frequently been shown to be sensitive to overexpression of nuclear RNaseH1. We, therefore, generated a U2OS cell line, which overexpresses the nuclear version of human RNaseH1 (M27-RNaseH1) under the control of a doxycycline-inducible promoter. Importantly, overexpression of nuclear RNaseH1 resulted in partially but significantly reduced DDX41 knockdown-dependent yH2AX and 53BP1 foci (Figure 9c, 9d). Moreover, neutral comet assay revealed that physical DSBs, which appear in the absence of DDX41, are suppressed by overexpression of M27-RNaseH1, suggesting that DDX41 opposes R-loop-dependent genomic instability (Figure 9e). The partial rescue of

spontaneous DNA damage after loss of DDX41 by RNaseH1 overexpression might be indicative of additional R-loop-independent DNA damage in DDX41 knockdown cells. However, overexpression of RNaseH1 has dramatic effects on cell proliferation, transcription, and DNA repair since also regulatory RNA-DNA hybrids are removed[385,386]. Thus, the partial rescue may also result from a dynamic situation, in which overexpressed RNaseH1 hydrolyzes deleterious DDX41 knockdown-dependent R-loops, while also removing native R-loops. Overexpression of RNaseH1, therefore, suppresses DNA damage at sites caused by depletion of DDX41, while potentially generating DNA damage in other regions. Comparable rescue of R-loop-dependent DNA damage has been observed in other studies. Expression of pathogenic MDS mutants of SF3B1 in K562 cells resulted in R-loop accumulation and spontaneous yH2AX formation, which was only partially suppressed by RNaseH1 overexpression[387]. In addition, increased R-loop-dependent pATM levels in Werner syndrome cells were only partially rescued by overexpression of RNaseH1[275].

We performed sBLISS in DDX41 knockdown cells to map the genome-wide distribution of DSB hotspots. As previously reported, promoter regions displayed spontaneous DNA fragility hotspots in unperturbed control cells[354,355]. Elevated accumulation of spontaneous DSBs in the promoter region upon depletion of DDX41 points to a function in opposing transcription-dependent genomic instability. Especially the promoter region is prone to endogenous DNA fragility through conflicts between replication and transcription machineries. To test if DNA damage in DDX41 knockdown cells was dependent on replication, we monitored spontaneous yH2AX levels throughout different cell cycle stages. Cell cycle-dependent yH2AX analysis in DDX41 knockdown cells revealed that spontaneous DNA damage peaked during the S-phase of the cell cycle, suggesting replication-dependent DNA damage induction (Figure 10a).

3.4 DDX41 knockdown cells suffer from DNA replication stress

Although the strongest yH2AX activation was evident during S-phase, we observed significantly increased levels throughout the cell cycle. Deleterious R-loops in DDX41 knockdown cells might be processed by NER factors, thereby inducing R-loop-dependent DNA damage independent of the cell cycle stage[207]. However, we speculate that DNA damage upon DDX41 depletion arises during S-phase through TRCs and is propagated into G2/M and subsequent G1 phase since yH2AX levels peaked during S-phase. Accumulation of TRCs during the S-phase caused by excessive R-loop formation has been reported as a frequent driver of replication stress and spontaneous DNA

damage[219,366,370,388,389]. Accordingly, DDX41 knockdown cells displayed significantly higher levels of the ATR-dependent replication stress marker pRPA (Ser33) (Figure 10b). Moreover, progression of replication forks was significantly impaired to a similar extent as in cells treated with a low dose of the DNA polymerase inhibitor aphidicolin (Figure 10c). We also observed significant alterations of Ser2/5 phosphorylation dynamics of the CTD of RNAPII, indicating perturbations during transcription initiation and elongation. Taken together, these findings suggest increased replication stress due to TRCs and activation of the replication stress response. In line, knockdown of DDX41 in U2OS cells or overexpression of pathogenic AML variants in OCI-AML 3 cells led to ATR inhibitor sensitivity. R-loop-dependent TRCs induce distinct DNA damage responses based on the orientation of the conflict[172]. Co-directional TRCs have been shown to induce the ATM-CHK2 axis, whereas HO conflicts result in activation of the ATR-CHK1 pathway[172]. Dependency of cells on ATR signaling and phosphorylation of prominent ATR substrates in the absence of DDX41 suggest that DDX41 knockdown cells primarily experience deleterious HO TRCs. Potential benefits from treatment with ATR inhibitors for MDS/AML patients with pathogenic DDX41 mutations should be further explored.

3.5 DDX41 regulates promoter-associated R-loops

The presence of DDX41 at R-loops under native conditions and spontaneous R-loop-dependent genomic instability in DDX41 knockdown cells prompted us to investigate the potential role of DDX41 as an R-loop resolvase. Indeed, S9.6 dot blot analysis revealed a global accumulation of R-loops in the absence of DDX41 (Figure 12d). Usage of the S9.6 antibody for R-loop detection is under strong debate since it also recognizes ssRNA and dsRNA species[362]. To rule out artifacts, we treated dot blot samples with RNaseIII and RNaseT1 to deplete ssRNA and dsRNA before spotting the nucleic acids on the nitrocellulose membrane. Moreover, we included an *in vitro* RNase H digestion control to show S9.6 specificity towards RNA:DNA hybrids. We complemented the dot blot analysis with an S9.6-independent HBD-GFP chromatin retention assay, which also revealed an accumulation of R-loops after DDX41 depletion (Figure 12e). Importantly, excessive R-loop formation was suppressed by overexpression of nuclear RNaseH1 or transcription inhibition using DRB. Global R-loop gain in the absence of DDX41 tempted us to hypothesize that DDX41 acts as an R-loop resolvase. We purified full-length DDX41 to test its *in vitro* binding and unwinding ability toward RNA-DNA hybrids. Before our study, very little was known about the potential substrates of DDX41 and its enzymatic activity. Previous structural analysis revealed a conserved binding mode to ATP, similar to the one identified in DDX3, underlining its role as an ATP-

dependent helicase[266]. Moreover, modeling of the DDX41 structure together with dsDNA suggests that binding to double-stranded nucleic acids is facilitated by the DEAD-domain and that the helicase domain is dispensable for binding[266]. Furthermore, DDX41 hydrolyzes ATP in the presence of different dsDNA substrates, indicative of enzymatic activity towards double-stranded nucleic acid substrates[390]. Our study provides the first evidence that purified DDX41 displays stronger binding towards RNA-DNA hybrids than dsDNA *in vitro* (Figure 12c). In accordance, DDX41 within nuclear extracts associates stronger with an RNA-DNA hybrid substrate in comparison to a dsDNA substrate[214]. Moreover, we found that DDX41 hydrolyzes ATP in the presence of RNA-DNA hybrids and displays unwinding activity towards RNA-DNA hybrid substrates (Figure 13d, 13e). Strikingly, a helicase domain-deficient DDX41 mutant failed to unwind RNA-DNA hybrids *in vitro* (Figure 13f). Various helicases with RNA-DNA hybrid unwinding activity have been reported, indicating that cells require multiple R-loop resolvases which regulate spatial and temporal distinct subsets of R-loops. SETX and DDX5 have been reported to unwind damage-induced RNA-DNA hybrids at DSBs[101,353]. The nucleopore-associated DDX19 helicase re-localizes to the nucleus after replication stress in an ATR-dependent manner to prevent R-loop-dependent replication fork stalling[217]. DDX39B/UAP56 ensures transcription elongation by unwinding R-loops along the gene body and the transcription termination site[219]. Similarly, DHX9 associates with transcription termination sites to support the release of RNA Pol II and faithfully terminate transcription[152]. In contrast, we demonstrated that DDX41 associates with promoter regions of highly transcribed genes *in vivo* to unwind promoter-proximal R-loops, thereby ensuring RNA Pol II elongation initiation. Loss of DDX41 resulted in excessive accumulation of promoter-associated R-loops in highly transcribed genes. Importantly, the role of DDX41 in maintaining promoter-proximal R-loop homeostasis was evident in U2OS cells and HCT116 cells, suggesting that DDX41 is a general promoter-associated R-loop resolvase in human cells. Further analysis revealed a striking overlap between promoters that accumulated R-loops and DSBs in DDX41 knockdown cells, indicating that DDX41 safeguards promoters against R-loop-dependent genomic instability. Intriguingly, we observed promoters containing R-loop gains which did not experience DNA fragility, suggesting that not all stabilized R-loops in the absence of DDX41 are translated into DNA damage. Comparison between genomic regions that are prone to DNA fragility after R-loop stabilization and regions that do not accumulate DSBs might reveal mechanistically why some R-loops lead to DNA damage while others do not. As previously described[172], we speculate that the orientation between the replication fork and the R-loop-

associated transcription machinery might be decisive whether DSBs arise after excessive R-loop formation in the promoter region. Moreover, overlap of R-loops and replication origins at CGIs might contribute to R-loop-dependent fragility. Future OK-sequencing and ORC1 CUT&RUN-sequencing analysis after loss of DDX41 are required to map directionality of the replication fork and sites of replication origins genome-wide. Correlation between our MapR/sBLISS analysis, replication fork directionality and sites of replication origins might elucidate why a specific subset of R-loops leads to DNA damage after stabilization. Moreover, we observed DSB hotspots in DDX41 knockdown cells, which were independent of R-loop gain. Intriguingly, we identified significant changes in the transcriptome of DDX41 knockdown U2OS cells by RNA-sequencing. Transcriptomic changes in genes that promote genomic stability after loss of DDX41 might cause secondary DSB formation at sites without R-loop gain in addition to primary DSB formation at sites with excessive R-loop formation. A more detailed view of individual changes in the transcriptome, complemented by proteome analysis of DDX41 knockdown cells is required.

3.6 Pathogenic DDX41 mutations predispose to familial MDS/AML

Pathogenic germline and somatic mutations in *DDX41* are frequently found in patients with MDS/AML[269,270,360]. Primarily, germline mutations appear as frameshifts, resulting in a premature protein[265]. In addition, mutations in the DEAD and the helicase domains were identified, hinting toward loss of function mutations. Mutations in major splicing regulators such as SRSF2, SF3B1, U2AF1, and ZRSR2 define a predominant group of mutations associated with MDS/AML[210,212]. Due to the high frequency of splicing factor mutations in MDS/AML patients and the identification of pathogenic DDX41 variants, previous studies speculated about a potential role of DDX41 in splicing[267,269,270]. Indeed, acute depletion of the *C.elegans* orthologue of human DDX41, SACY-1 results in altered 3´splice site usage and transcriptome changes[262]. Similarly, loss of human DDX41 leads to pre-mRNA splicing alterations and differential RNA processing[270]. Moreover, co-immunoprecipitation studies of endogenous DDX41 revealed an association with multiple splicing complexes such as the A, B, U2, U4/U6, and PRPF19 complex[270]. Due to the unspecific broad interaction with nearly all splicing complexes, we speculate that DDX41 is not an active member of the spliceosome, but rather associates through its high occupancy of promoter regions, at which the spliceosome is assembled onto the nascent transcript[391]. However, we hypothesize that splicing and transcriptome alterations are a consequence of excessive promoter-proximal R-loop formation in the absence of DDX41. Correlation between RNA Pol II transcription speed and co-transcriptional splicing has been reported[66,68,392]. Accordingly, we speculate that deficient release

of RNA Pol II from promoter-proximal pausing due to aberrant R-loop unwinding in the absence of DDX41 leads to splicing alterations and transcriptomic changes. Interestingly, a recent study suggests that Ddx41 deficiency in zebrafish causes deleterious R-loop accumulation in HSPCs, upregulated inflammatory signaling, and aberrant expansion of HSPCs[260]. Moreover, augmented R-loop formation caused by pathogenic splicing factors mutations has recently been proposed as a common driver for MDS/AML, instead of splicing alterations itself[210]. We demonstrate that knockdown of DDX41 leads to promoter-associated R-loop accumulation and to spontaneous DSB formation in human HSPCs (Figure 19a). Similarly, even though less striking than the knockdown, expression of the AML-associated pathogenic DDX41 R525H mutant in HSPCs caused DSB induction, whereas no DSB formation was observed when wild-type DDX41 was expressed (Figure 19b). *In vitro*, we observed a mild but significant reduction in unwinding activity towards RNA-DNA hybrids when comparing the R525H variant with wild-type DDX41 (Figure 13f). Since we observed elevated DNA damage after exogenous expression of DDX41 R525H in the presence of the endogenous wild type, we speculate that the R525H variant acts as a dominant-negative mutant, which displaces the wild type from the promoter region, while not being able to faithfully unwind promoter-associated R-loops. Thus, the expression of R525H in HSPCs might lead to excessive R-loop accumulation and spontaneous R-loop-dependent DSBs, thereby contributing to the development of MDS/AML.

Promoter-associated R-loops have been reported to regulate transcriptional dynamics[73]. Loss of promoter-proximal R-loop homeostasis in the absence of DDX41 potentially results in transcriptomic changes. Intriguingly, DDX41 knockdown cells displayed R-loop accumulation within the promoter region of genes, which are frequently altered in AML such as chromatin organization, RUNX1 interaction as well as NOTCH and TGF-β signaling (Figure 15h). Multiple driver mutations in genes coding proteins involved in chromatin organization have been identified in AML patients. Prominent mutations are found in *TET2* and *DNMT3A/B*, which regulate DNA methylation dynamics, thereby influencing chromatin compaction and transcription[393–397]. Accordingly, targeting DNA methylation is being explored as a promising therapeutic strategy for AML patients[398]. Interestingly, demethylation therapy of AML patients with pathogenic *DDX41* mutations displays an overall high response rate of 69%[399]. Furthermore, mutations in members of the cohesion complex such as *STAG2* and mutations in chromatin accessibility factors such as *HMGN1* are associated with MDS/AML[400–402]. In addition, genes contributing to chromatin

organization, AML driver mutations in the runt-related transcription factor 1 (RUNX1) have been identified[360,403,404]. Moreover, upregulated TGF-β and NOTCH signaling have been linked to the progression of AML[405–407]. Especially deregulated TGF-β signaling in mesenchymal stromal cells within the micro-environment of HSPCs has been described as a relevant cause for MDS/AML[405]. Transcriptomic changes in these pathways caused by aberrant promoter-proximal R-loop processing in the presence of pathogenic DDX41 mutations may contribute to MDS/AML development and progression.

As mentioned above, Ddx41-deficient zebrafish HSPCs displayed an active inflammatory response and altered HSPC expansion[260]. We demonstrated that loss of DDX41 results in spontaneous DNA damage, which is accompanied by upregulation of the inflammatory response and activation of the NF-kB pathway. Two distinct mechanisms may induce inflammation triggered by R-loop-dependent DNA damage in DDX41 knockdown cells: (1) R-loops are identified and excised by the nucleotide excision repair proteins XPC/XPG and subsequently exported from the nucleus to the cytoplasm; (2) R-loop-dependent TRCs cause DNA damage and replication stress that both induce aberrations during mitosis. Chromosome segregation errors result in micronuclei formation in the subsequent G1 phase, during which nucleic acids might leak into the cytoplasm due to micronuclei rupture. In both scenarios, cytoplasmic nucleic acids would subsequently be detected by cGAS and activate the cGAS-STING pathway, which promotes the induction of inflammatory transcription factors such as IRF3 and NF-kB. Inflammatory signaling in HSPCs and their surrounding micro-environment has been reported to influence HSPC activation, expansion, and myeloid differentiation[304,307,310]. Alterations during hematopoiesis including expansion of HSPCs and aberrant differentiation define classical features of MDS and AML. In accordance, DDX41 knockout mouse embryos display dramatic defects in HSPC expansion and HSPC differentiation, confirming that DDX41 is a major regulator of hematopoietic development[408]. We propose that pathogenic DDX41 mutants in MDS/AML patients contribute to disease development and progression through two distinct mechanisms: (1) malignant HSPC expansion and differentiation through excessive R-loop and DSB accumulation in the promoter region as well as DNA-damage-induced inflammatory signaling; (2) transcriptomic changes in genes involved in chromatin organization, RUNX1 regulation as well as NOTCH and TGF-ß signaling, caused by aberrant unwinding of promoter-proximal R-loops. Finally, the dependency of DDX41 knockdown cells on

ATR signaling suggests exploring ATR inhibition as a therapeutic approach in MDS/AML patients with *DDX41* mutations.

3.7 Genuine R-loop formation

R-loops have been reported to form at distinct genomic loci, such as centromeres[95,112], telomeres[92], transcription start sites[156,255], gene bodies[151], and transcription termination sites[151]. Striking discrepancies between different R-loop mapping approaches have raised doubts about the genuineness of different subsets of R-loops[72,409]. The initial hypothesis that different classes of R-loops are identified by RnaseH and S9.6-based approaches has recently been refuted[256]. Instead, short promoter-proximal R-loops (class I) are predominantly identified by native mapping strategies, while *ex vivo*-based approaches mostly recover longer R-loops that form throughout the gene body and the TTS (class II)[72,256]. So far, no convincing argument has been made why the two different types of protocols recover distinct R-loop classes. Although both classes could co-exist, there is also the possibility that one of the experimental pipelines identifies artefactual R-loops while the other identifies genuine R-loops. Strikingly, gene body R-loops which are mostly identified by RNA-based sequencing approaches such as DRIPc-seq and RR-ChIP, significantly overlap with RNAs derived from repeats such as Alu and LINE elements[409]. Although these RNA species could hypothetically form R-loops *in trans*[410], it is more likely that these are artefactual R-loops that result from co-purified repeat-containing dsRNA species, since they are not identified by DNA-based and native R-loop mapping approaches. Moreover, gene body R-loop signal identified by DRIPc-seq in fission yeast is sensitive to RNase III treatment, suggesting to result from dsRNA[411]. Unlike R-loops in gene bodies, TTS-associated R-loops are also recovered by DNA-based *ex vivo* sequencing approaches, suggesting that they are genuine class II R-loops[152,389]. Promoter-proximal class I R-loops are identified by all types of DNA-based R-loop sequencings, thus representing the second genuine type of R-loops[156,255,389]. High GC-skew in promoter regions further supports the idea of genuine promoter-proximal R-loops. Stabilization of co-transcriptional R-loops would require pausing of RNA Pol II, which is unlikely to happen within gene bodies under unperturbed conditions to an extent observed by DRIPc-seq and RR-ChIP. In contrast, promoter-proximal pausing of RNA Pol II is a well-established event, supporting the idea of promoter-proximal R-loops[52]. Moreover, RPA binding on both DNA strands and significantly reduced TOP1 activity at TSSs suggest that promoter-proximal R-loops constitute the predominant class of genuine R-loops[409].

It remains elusive which subset of R-loops drives genomic instability. Although many studies have observed a global increase in R-loops and DNA damage signaling, only a few studies have mechanistically investigated which genomic loci are prone to R-loop-dependent DNA damage. Endogenous DNA fragility hotspots are mainly mapped to gene promoters which are likely to be the predominant site of genuine R-loop formation[354,355]. Thus, it is tempting to speculate that mostly promoter-proximal R-loops contribute to spontaneous DSB formation. In accordance, we found significantly increased spontaneous DNA fragility hotspots at gene promoters that experienced excessive R-loop accumulation in DDX41 knockdown cells. While endogenous DSBs predominantly occur at gene promoters, R-loop-dependent DNA fragility has also been observed at TTSs after cellular perturbations. TOP1 knockdown cells display R-loop stabilization, pRPA binding, and DSBs at TTSs[389], suggesting that in addition to TSS-associated R-loops also accumulation of R-loops at TTSs can induce DNA fragility. Not all genomic sites that displayed R-loop stabilization upon knockdown of DDX41 or TOP1 also experienced DNA fragility. This observation raises the question of which genomic features determine whether R-loops are translated into DNA breaks. A key determinant appears to be the directionality of the conflict between the progressing replication fork and the R-loop-associated transcription machinery. HO TRCs predominantly lead to further R-loop formation, while CD TRCs rather promote R-loop resolution. Accordingly, in TOP1 knockdown cells, HO-orientated conflicts result in increased R-loop formation at TTSs compared to CD-orientated conflicts[389]. Thus, replication fork polarity likely constitutes a crucial genomic feature that contributes to the decision of whether R-loop formation becomes detrimental to genomic integrity. Promoter regions of active genes which are highly occupied by RNA Pol II often contain replication origins[188,412]. Therefore, R-loop accumulation in promoters that also serve as replication origins may be more deleterious than in promoters that do not contain origins. Moreover, NER structure-specific nucleases XPG and XPF actively process R-loops into DSBs[207]. Genomic properties that promote the recruitment of the NER machinery, such as gene transcription rates and chromatin accessibility may display TRC-independent determinants of whether R-loops become toxic to the genome[413]. I speculate that genuine R-loops form at TSSs and TTSs and that dysregulation of both types of R-loops can lead to DNA fragility, while mainly TSS-associated R-loops contribute to endogenous DNA damage. Moreover, I hypothesize that the orientation of the replication fork and the R-loop-associated transcription machinery, presence of replication origins within gene promoters, gene transcription rates, and

chromatin accessibility constitute the key determinants of whether dysregulated R-loops induce genomic instability.

4 Material and Methods

4.1 Lists of consumables, equipment, and software

General chemicals and solutions	
Item	**Supplier/composition**
Acetic acid	Sigma Aldrich
Acetone	Sigma Aldrich
Acetonitrile (ACN)	Sigma Aldrich
ATP	Sigma Aldrich
EDTA	Sigma Aldrich
EGTA	Sigma Aldrich
Ethanol	Sigma Aldrich
Formic acid	Sigma Aldrich
Glycerol	Sigma Aldrich
HEPES	Sigma Aldrich
Methanol	Sigma Aldrich
Manganese chloride ($MnCl_{2)}$	Sigma Aldrich
Potassium chloride	Sigma Aldrich
Sodium ascorbate	Sigma Aldrich
Sodium azide	Sigma Aldrich
Sodium chloride (NaCl)	Sigma Aldrich
Sodium deoxychylate	Sigma Aldrich
Trifluoroacetic acid (TFA)	Sigma Aldrich
Triton-X-100	Sigma Aldrich
Urea	Sigma Aldrich

Cell culture	
0.05% Trypsin-EDTA	Gibco
Fetal bovine serum (FBS)	Gibco
100 U/ml Penicillin/Streptomycin	Gibco
L-glutamine	Gibco
Alpha MEM Eagle, without L-Glutamine	PAN-Biotech
Amaxa® Human CD34+ Cell Nucleofector® Kit	Lonza

ATR inhibitor VE-821	Selleckchem
Biotin-phenol	Iris Biochem
Dialyzed FBS (10,000 molecular weight cut-off)	Sigma Aldrich
Dimethyl sulfoxide (DMSO)	Sigma Aldrich
Doxycycline (DOX)	Sigma Aldrich
D-MEM for SILAC without lysine and arginine	Life Technologies
Dulbecco's Modified Eagle Medium (D-MEM)	Gibco
Dulbecco's Phosphate-Buffered Saline (D-PBS)	Gibco
Fetal bovine serum (FBS)	Gibco
G418	Invivogen
G-CSF	StemCell Technologies
GM-CSF	StemCell Technologies
Human bone osteosarcoma epithelial cells (U2OS)	ATCC
Human HCT116	ATCC
Human CD34+ cells	Lonza
Human embryonic kidney cells (HEK293/T)	ATCC
Human OCI-AML3	DSMZ
Hydrogen peroxide	Sigma Aldrich
StemSpan SFEM II	StemCell Technologies
TPO	StemCell Technologies
L-arginine (Arg0)	Cambridge Isotope Laboratories
L-arginine- U-13C6 99% (Arg6)	Cambridge Isotope Laboratories
L-arginine-U-13C6-15N4 99% (Arg10)	Cambridge Isotope Laboratories
L-lysine (Lys0)	Cambridge Isotope Laboratories
L-lysine- U-13C6-15N2 99% (Lys8)	Cambridge Isotope Laboratories
L-lysine-4,4,5,5,-D4 96–98% (Lys4)	Cambridge Isotope Laboratories
Linear polyethylenimine transfection (PEI, HCl Max, 40000)	Polysciences, Inc.
Lipofectamine RNAiMAX	Life Technologies
Puromycin	Invivogen
Opti-MEM with GlutaMAX	Gibco
RLT buffer	Qiagen

Cell lysis	
Complete protease inhibitor cocktail tablets	Roche Diagnostics

Dithiothreitol (DTT)	Sigma Aldrich
Modified RIPA buffer	50 mM Tris-HCl pH 7.5, 150 mM NaCl, 1 mM EDTA, 1% NP-40, 0.1% Sodium-deoxycholate
N-ethylmaleimide (NEM)	Sigma Aldrich
NuPAGE LDS Sample Buffer (4×) (LDS SB)	Thermo Fisher Scientific
Phosphatase inhibitors:	Sigma Aldrich
Sodium orthovanadate	Sigma Aldrich
β-glycerophosphate	Sigma Aldrich
Sodium fluoride	Sigma Aldrich
Sodium dodecyl sulfate	BioRad
QuickStart Bradford 1x Dye Protein Reagent	50 mM Tris-HCl, pH 7.4, 1% Triton X-100, 0.5%
RIPA buffer	Sodium-deoxycholate, 0.1% SDS, 150mM NaCl
SDS-PAGE / Western blotting	
0.45 µm nitrocellulose	Sigma Aldrich
Blocking buffer	10% skimmed milk solution in PBS-T
Bovine serum albumin (BSA)	Sigma Aldrich
NuPAGE MOPS SDS running buffer (20X)	Thermo Fisher Scientific
NuPAGE Bis-Tris gels 4-12%	Thermo Fisher Scientific
PBS-T	1x PBS, 0.1% Tween-20
Primary antibody solution	5% BSA in PBS-T, 0.06% sodium azide
SuperSignal West Pico Chem. Substrate	Thermo Fisher Scientific
Transfer buffer	25 mM Tris, 192 mM Glycine, 20% (v/v) methanol, pH 8.3
Immunofluorescence microscopy	
4% paraformaldehyde (PFA) in PBS	Affymetrix
Alexa Fluor 647 azide	Thermo Fisher Scientific
Blocking Buffer	3% BSA in PBS-T
Bovine serum albumin (BSA)	Sigma Aldrich
Hoechst 33342	Thermo Fisher Scientific
Vectorshield	Vector Laboratories
Cloning	
BL21 DE3 codon+	Agilent
Dh5α	Thermo Fisher Scientific
Dpn1	New England Biolabs

Gateway LR Clonase II Enzyme Mix	Invitrogen
High-efficiency DH5α E.coli	New England Biolabs
NucleoBond Xtra Midi	Macherey-Nagel
Plasmid Mini Kit	Qiagen
Q5 MasterMix	New England Biolabs
QIAquick PCR & Gel Cleanup Kit	Qiagen
In-gel digestion	
Buffer B	80% ACN, 0.5% acetic acid
Chloroacetamide (CAA)	Sigma Aldrich
Colloidal Blue Staining Kit	Life Technologies
Destaining solution	50% Ethanol, 50 mM ABC buffer pH 8.0 25 mM
Digestion buffer	ABC buffer pH 8.0
Peptide extraction buffer	30% ACN, 3% TFA
Sequencing grade Trypsin (0.5 µg/µl in 50 mM acetic acid)	Sigma Aldrich
Stage tipping	
Buffer A	0.1% formic acid
Buffer A*	5% ACN, 0.1% TFA
Buffer B	80% ACN, 0.1% formic acid
C_{18} elution buffer	50% ACN, 0.1% formic acid
C_{18} Empore 47 mm extraction disks	CDS Analytical
Dot blot analysis	
β-mercaptoethanol	Sigma Aldrich
DNeasy mini kit	Qiagen
Nitrocellulose membrane	NeoLab Migge
RNaseIII	Produced in-house by Martin Möckel
RnaseH	New England Biolabs
RNase T1	Thermo Fisher Scientific
RNA sequencing	
DNase	New England Biolabs
RNeasy Plus Mini Kit	Qiagen
RNase-free water	Sigma Aldrich
RNAaseZAP	Sigma Aldrich
TruSeq Stranded mRNA LT Sample Prep Kit	Illumina

MapR	
2× Stop Buffer	68 µl 5M NaCl, 40 µl 0.5 M EDTA, 20 µl 0.2 M EGTA, 10 µl 5% digitonin, 5 µl 10mg/ml RNaseA
ActinomycinD	Cell Signaling Technology
Binding buffer	20 mM Hepes-KOH pH 7.9, 10 mM KCl, 1mM CaCl2, 1 mM MnCl2
Concanavalin A-coated beads	Polysciences Europe
dsDNA HS Assay Kit	Qubit
Digitonin	Millipore
Proteinase K	New England Biolabs
Spermidine	Sigma Aldrich
Ovation Ultralow System V2	NuGEN
Wash buffer	Hepes-NaOH pH7.5, 150 mM NaCl, 0.5 mM Spermidine, 1 mM protease inhibitor
RDProx	
NeutrAvidin agarose beads	Thermo Fisher Scientific

Commercial assays	
Comet assay	Trevigen
Comet slides	Trevigen
In Situ PLA	Duolink
ADP-Glo assay	Promega
Cell titer blue viability assay	Promega
sBLISS	
Klenow fragment	New England Biolabs
NEBNext dA-tailing	New England Biolabs
CUTSmart buffer	New England Biolabs
greenCUT&RUN	
37% Formaldehyde	Covaris
Glycine	Sigma Aldrich
Accel-NGS 1S Plus DNA library kit	Swift Biosciences
Multiplex oligos, dual indexing	Illumina
Zymoclean gel DNA recovery kit	Zymo
DNA fiber assay	
5-Chloro-2′-deoxyuridine	Sigma Aldrich

5-Iodo-2′-deoxyuridine	Sigma Aldrich
Aphidicolin	Sigma Aldrich
HCl	Sigma Aldrich
Prolong gold AntiFade mountant	Thermo Scientific Fisher
SuperFrost+ microscopy slides	Thermo Scientific Fisher

Protein production	
Amicon spin concentrators	Merck Millipore
GSTrap HP 5 ml	Cytiva
HisTrap FF crude	Cytiva
HisTrap FF 5ml	Cytiva
HisTrap Heparin HP 5 ml	Cytiva
His6-3C protease	New England Biolabs
Imidazole	Sigma Aldrich
Polyethylenimine	Sigma Aldrich
SF9000 II media	Thermo Scientific Fisher
Superdex 200 16/60 pg	Cytiva
Superdex 75 16/60 pg	Cytiva

Antibodies	**Product number**	**Origin**	**Dilution (WB/IF)**
γH2AX	A300-081A-M	Bethyl	(1:1000/1:500)
53BP1	MAB3802	Millipore	(-/1:200)
pRPA (S33)	A300-246A-M	Bethyl	(-/1:200)
DDX41	15076	Cell Signaling	1:1000
GFP	sc-9996	Santa Cruz	1:1000
S9.6	ENH001	Kerafast	1:10000
dsDNA	ab27156	Abcam	1:1000
AQR	A302-547A	Bethyl	1:2000
DDX42	SAB1407136	Sigma	1:1000
DDX39A	SAB2700315	Sigma	1:1000
FLAG M2	F1804	Sigma	1:2000
B-Actin	A2228	Sigma	1:10000
DHX37	A300-856A-M	Bethyl	1:1000
BrdU (mouse)	347580	BD Bioscience	1:100
BrdU(rat)	ab6326	Abcam	1:500
Anti-mouse Cy3.5	Ab6946	Abcam	1:100
Anti-rat Cy5	Ab6565	Abcam	1:100

P65	sc-372	Santa Cruz	(-/1:200)
Phospho-Rpb1-CTD (Ser2)	13499	Cell signaling	(-/1:200)
Phospho-Rpb1-CTD (Ser5)	13523	Cell signaling	(-/1:200)

Gene name	**Sequence 5'-3' or origin**
AQR	CUGAAUAUGGCGGUGUAGU
DDX41	L-010394-00-0005 – Horizon Discovery
DDX42	L-012393-01-0005 – Horizon Discovery
DHX37	L-019073-00-0005– Horizon Discovery
DDX39A	L-004920-01-0005 – Horizon Discovery
DDX27	L-013635-01-0005 – Horizon Discovery
Si control pool	D-001820-10 – Horizon Discovery

Plasmid	**Origin**
pcDNA-DEST53	Life Technologies
plix402	Addgene
pDEST53-APEX2-FLAG-HBD	This study
pDEST53-APEX2-FLAG-HBD-WKK	This study
pDEST53-GFP-HBD	This study
pDEST53-GFP-HBD-WKK	This study
plix402-GFP-DDX41-WT	This study
plix402-GFP-DDX41-L237FP238T	This study
plix402-GFP-DDX41-R525H	This study
His6-GST-3CDDX41-WT	This study
His6-GST-3C-DDX41-R525H	This study
His6-GST-3C-HBD-AVI	This study
His6-GST-3C-HBD-WKK-AVI	This study
pET21b-GFPnanobody(LaG16)-MNase-His6	This study

Oligonucleotide	**Sequence 5'-3'**
DDX41-fw-qPCR	GTCCGTGAAAGAGCAGATGGAG
DDX41-rev-qPCR	GTAGCGACAGATGTCTAGGCTG
HBD-backbone-fw	cctcctcacggcatagaacatggtggagcctgctttttgtacaaagttgg
HBD-backbone-rev	cctttgtcaggaaatctgcaagcgacccagctttcttgtacaaagt

HBD-insert-fw	ccaactttgtacaaaaaagcaggctccaccatgttctatgccgtgaggagg
HBD-insert rev	actttgtacaagaaagctgggtcgcttgcagatttcctgacaaagg
HBD-W43A-fw	ctttctgaccgcgaatgagtgcagagcacaggtggaccg
HBD-W43A-rev	accccggtcttgcggccc
HBD-K59A-K60A-fw	tgccagatttgcggcgtttgccacagaggatgaggc
HBD-K59A-K60A-rev	gcaggaaaccggtccacc
DDX41-R525H-fw	CGCACCGGGCACTCGGGAAAC
DDX41-R525H_rev	GCCAATCCGGTGTACATAGTTCTC
DDX41-L237F-P238T-fw	TGTTCACGTTTACCGTCATCATG
DDX41-L237FP238T-rev	CCAGTGTCTTGCCTGAAC
DDX41-3C-rev	ACCGGGCCCCTGGAACAGAAC
DDX41-L153-fw	TTCTGTTCCAGGGGCCCGGTCTGAGCATGTCTGAAGAGC
DDX41-Q410-rev	gtgctcgagtgcggccgctcactggatgacatccaggctg
DNA-12mer-fw	GACACCTGATTC-6-FAM
DNA-12mer-rev	GAATCAGGTGTC
RNA-12mer-fw	GACACCTGATTC-6-FAM
RNA-12mer-rev	GAATCAGGTGTC
DNA-38mer-IBFQ	TAAAACAAAACAAAACAAAACAAAATCTTTACGGTGCT
RNA-13mer-6-FAM	6-FAM-AGCACCGUAAAGA
DNA-38mer	TAAAACAAAACAAAACAAAACAAAATCTTTACGGTGCT

Software	**Version**
RStudio	1.3.959
STRING database	11.0b
Cytoscape	3.2.1
Reactome	V2021
PANTHER	16.0
EnrichR	Update 2016
CometScore Tritek	2.0
Fiji/ImageJ	1.51
MaxQuant	1.5.2.8
GraphPad Prism	7.04
Blc2fastq	2.19
Bowtie2	2.3.4
MACS2	2.1.2
Diffbind	3.0.5

ClusterProfiler	3.18
DeepTools	3.4.1
CHIPSeeker	1.26
STAR	2.7
featureCounts	1.6
Bioconductor	2.46
DESeq2	1.26
Harmony High-Content Imaging and analysis software (PerkinElmer)	4.4
EdgeR	3.32.1
BWA-MEM	0.7.15
Bedtools	2.27.0
Cutadapt	1.18
FastQC	0.1.2
CHIPpeakanno	3.28.0
GenomicRanges	1.46.9
Rtracklayer	1.54
GenomicFeatures	1.46.1

Machine	**Company**
3311 Forma Steri-Cult CO_2 Incubator	Eppendorf
AF7000	Leica
BD LSRFortessa SORP	BD Biosciences
Biometra TRIO Thermal Cyclers	Analytikjena
Biorupter NGS	Diagenode
Branson Sonifier 450	Thermo Fisher Scientific
ChemiDoc imaging system	BioRad
DMi8 inverted microscope	Leica
Dot blot apparatus	BioRad
Easy-LC-1000	Thermo Fisher Scientific
FACSMelody cell sorter	BD Biosciences
NanoDrop 2000	Thermo Fisher Scientific
NextSeq500	Illumina

NGC Quest Plus FPLC system	BioRad
NuPage Novex Gel System	Thermo Fisher Scientific
Opera Phenix High Content Screening System	PerkinElmer
Plate reader infinite m200	Tecan
Plate reader m20	Tecan
Q Exactive Plus	Thermo Fisher Scientific
Qubit 2.0 Fluorometer	Life Technologies
SP5 confocal microscope	Leica
SPE confocal microscope	Leica
Tapestation 2200	Agilent
Thermoshaker	Eppendorf
Typhoon FLA 9000	GE Healthcare
Vacufuge Plus	Eppendorf
Visiscope 5-elements spinning disc	Visitron Systems

4.2 Methods

4.2.1 Cell culture

U2OS, HCT116, and HEK293T cells were obtained from ATCC and cultured in D-MEM medium (U2OS and HEK293T) or RPMI 1640 (HCT116) supplemented with 10% fetal bovine serum, L-glutamine, penicillin, and streptomycin. OCI-AML3 cells were purchased from DSMZ GmbH and cultured in D-MEM medium (PAN-Biotech) containing 20% FBS, L-glutamine, penicillin, and streptomycin. Cells were routinely tested for mycoplasma infection with a PCR-based method. For SILAC labeling, cells were cultured in media containing either L-arginine and L-lysine, L-arginine [13C6], and L-lysine [2H4] or L-arginine [13C615N4] and L-lysine [13C6-15N2] (Cambridge Isotope Laboratories)[414]. All cells were cultured at 37°C in a humidified incubator containing 5% CO2.

4.2.2 MS analysis

Peptide fractions were analyzed on a quadrupole Orbitrap mass spectrometer (Q Exactive or Q Exactive Plus, Thermo Scientific) equipped with a UHPLC system (EASY-nLC 1000, Thermo Scientific) as described[415,416]. Peptide samples were loaded onto C18 reversed-phase columns (15 cm length, 75 µm inner diameter, and 1.9 µm bead size) and eluted with a linear gradient from 8 to 40% acetonitrile containing 0.1% formic acid in 2 h. The mass spectrometer was operated in data-dependent mode, automatically switching between MS and MS2 acquisition. Survey full-scan

MS spectra (m/z 300 – 1700) were acquired in the Orbitrap. The 10 most intense ions were sequentially isolated and fragmented by higher-energy C-trap dissociation (HCD)[417]. An ion selection threshold of 5,000 was used. Peptides with unassigned charge states, as well as with charge states less than +2 were excluded from fragmentation. Fragment spectra were acquired in the Orbitrap mass analyzer.

4.2.3 Peptide identification

Raw data files were analyzed using MaxQuant (development version 1.5.2.8)[347]. Parent ion and MS2 spectra were searched against a database containing 98,566 human protein sequences obtained from the UniProtKB released in 04/2018 using Andromeda search engine[418]. Spectra were searched with a mass tolerance of 6 ppm in MS mode, 20 ppm in HCD MS2 mode, strict trypsin specificity, and allowing up to 3 missed cleavages. Cysteine carbamidomethylation was searched as a fixed modification, whereas protein N-terminal acetylation and methionine oxidation were searched as variable modifications. The dataset was filtered based on posterior error probability (PEP) to arrive at a false discovery rate of below 1% estimated using a target-decoy approach[419].

4.2.4 RDProx

SILAC-labeled cells were transfected with a construct expressing APEX2-tagged HBD or WKK. After 48 hours, cells were pre-treated with 500 µM biotin phenol (Iris Biochem) for 2 hours at 37°C, followed by a 2-minute incubation with 1 mM H_2O_2 (Sigma-Aldrich) at room temperature. Cells were washed twice with quenching solution (10 mM sodium azide, 10 mM sodium ascorbate, 5 mM Trolox (all from Sigma-Aldrich), and twice with PBS. Cells were lysed on ice using RIPA buffer (50 mM Tris, 150 mM NaCl, 0.1% SDS, 0.5% sodium deoxycholate, 1% Triton X-100). To release chromatin-bound proteins, cell lysates were sonicated using Biorupter (Diagenode). For affinity purification of biotinylated proteins, equal amounts of differentially SILAC-labeled cell extracts, originating from either the HBD or the WKK condition, were combined before the pull-down and incubated with pre-equilibrated NeutrAvidin agarose beads (Thermo Scientific) for 2 hours at 4°C on a rotation wheel. Beads were washed once with RIPA buffer, thrice with 8 M Urea (Sigma) in 1% SDS, and once with 1% SDS in PBS. Bound proteins were eluted in NuPAGE LDS Sample Buffer (Life Technologies) supplemented with 1 mM DTT and boiled at 95°C for 15 min. The eluates, after cooling down to room temperature, were alkylated by incubating with 5.5 mM chloroacetamide for 30 min in the dark and then loaded onto 4-12% gradient SDS-PAGE gels. Proteins were stained using the Colloidal Blue Staining Kit (Life Technologies) and digested in-

gel using trypsin. Peptides were extracted from the gel and desalted on reversed-phase C18 Stage Tips.

4.2.5 RDProx network analysis

Pearson correlations were calculated using RStudio (version 1.3.959). Functional protein interaction network analysis was performed using interaction data from the STRING database[420]. Only interactions with a score > 0.7 are represented in the networks. Cytoscape (version 3.2.1) was used for visualization of protein interaction networks[421]. Genes were manually annotated by literature research and clustered based on similarity. PFAM domain enrichment analysis was performed using EnrichR[422]. The respective terms with the lowest FDR based on Fisher's exact test and correction for multiple comparisons are highlighted next to each cluster.

4.2.6 SDS-PAGE and Western blotting

Proteins were resolved on 4-12% gradient SDS-PAGE gels (NuPAGE® Bis-Tris Precast Gels, Life Technologies) and transferred onto nitrocellulose membranes. Membranes were blocked using 10% skimmed milk solution in PBS supplemented with 0.1% Tween-20. The list of antibodies used in this study and conditions can be found in section 4.1. Secondary antibodies coupled to horseradish peroxidase (Jackson ImmunoResearch Laboratories) were used for immunodetection. The detection was performed with SuperSignal West Pico Chemiluminescent Substrate (Thermo Scientific).

4.2.7 Neutral comet assay

Neutral comet assay was performed according to the manufacturer's protocol (Trevigen). Briefly, cells were embedded in low melting agarose at 37°C on Comet Slides (Trevigen). Overnight cell lysis at 4°C was followed by equilibration in 1× Neutral Electrophoresis Buffer for 30 min at room temperature. Single-cell electrophoresis was performed at 4°C in 1× Neutral Electrophoresis buffer for 45 min with constant 21V. After DNA precipitation with 1× DNA Precipitation Buffer, Comet Slides were dried with 70% EtOH at room temperature. To completely dry the samples, Comet Slides were transferred to 37°C for 15 min. DNA was stained with SYBR Gold solution for 30 min at room temperature. Images were taken with a Leica AF7000 microscope using a 20× 0.8NA air objective and a filter cube 480/40 nm, 505nm, and 527/30 for excitation, dichroic, and emission wavelengths respectively. Tail moments of the comets were quantified using the CometScore (TriTek Corp.) software. At least 50 comets were quantified per condition.

4.2.8 RNA-DNA hybrid dot blot

Genomic DNA was extracted using the DNeasy mini kit (Qiagen). The isolated gDNA was treated with 1.2 U RNase III (produced in-house) for 2 h at 37°C. After enzyme deactivation at 65°C for 20 min, samples were split in half to digest control samples with 10 U RNaseH1 (NEB) overnight at 37°C. Enzyme deactivation was followed by spotting DNA in a serial dilution on a nitrocellulose membrane (NeoLab Migge GmbH) using a dot blot apparatus (BioRad). DNA was cross-linked to the membrane by UV light and afterward blocked with 10% skimmed milk solution in PBS supplemented with 0.1% Tween-20. The membrane was incubated overnight at 4°C with the S9.6 antibody (produced in-house). After incubation of secondary antibodies conjugated to horseradish peroxidase (Jackson ImmunoResearch Laboratories) signal was detected using SuperSignal West Pico Chemiluminescent Substrate (Thermo Scientific). An antibody against dsDNA was probed as a loading control after stripping the membrane with β-mercaptoethanol (Sigma) and 0.1% SDS in PBS. The detected signal was quantified using Fiji/ImageJ (v1.51) and ratios between the signal resulting from S9.6 and dsDNA staining were calculated to quantify global R-loop levels[423].

4.2.9 Proximity ligation assay

Proximity Ligation Assay was performed according to the manufacturer´s protocol (Duolink®, Sigma-Aldrich). Cells were fixed with 4% paraformaldehyde in PBS and permeabilized with 0.25% Triton X-100. Samples were blocked with Duolink® Blocking Solution for 1 hour at 37°C in a humidity chamber. After removal of the blocking solution, primary antibodies diluted in Duolink® Antibody Diluent were added to the coverslips for 2 hours at room temperature in a humidity chamber. Coverslips were washed 2× with Washing Buffer A. PLA plus and minus probes were put on in a 1:5 dilution in Duolink® Antibody Diluent for 1 hour at 37°C in a humidity chamber. Two washes with Washing Buffer A were followed by Ligase treatment in 1× Ligation Buffer for 30 min at 37°C in a humidity chamber. Ligation buffer was tapped off and coverslips were washed twice with Washing Buffer A. Amplification was achieved by adding the Polymerase in 1× Amplification buffer for 100 min at 37°C in a humidity chamber. After washing the samples 2× with 1× Washing Buffer B and 1× with 0.01× Washing Buffer B, coverslips were stained with 1 µg/ml Hoechst33342 and mounted using Dako mounting medium. Images were taken with a Leica SPE microscope using a 63× 1.4NA oil objective. The number of PLA spots per nucleus was quantified using Fiji/ImageJ (v1.51)[423].

4.2.10 ATPase assay

The ADP-Glo Assay was performed according to the manufacturer´s protocol (Promega). In brief, an ATP/ADP standard curve was prepared before each experiment to interpolate the measured values. Purified full-length DDX41 was incubated in a serial dilution together with 100 nM of RNA-DNA substrate with an ssDNA overhang and 5 µM ATP. After incubating the mix at 37°C for 60 min, the reaction was stopped by depleting unconsumed ATP with the ADP-Glo Reagent. The Kinase Detection Buffer was added to convert ADP to ATP and to add luciferase and luciferin to detect ATP. The resulting luminescence was measured with a Spark M200 (Tecan). The measured values were interpolated based on the values obtained by the ATP/ADP standard curve using GraphPad PRISM (v7.04, Graphpad Software, Inc.).

4.2.11 Fluorescence polarization assay

DsDNA, dsRNA, and RNA-DNA hybrid 12-mer substrates were generated by heating the respective 6-FAM-conjugated and unlabeled oligonucleotide pairs to 95°C and gradually cooling them down to 4°C. Single-stranded and double-stranded substrates were diluted to a final concentration of 20 nM in FP assay buffer (20mM HEPES, pH 7.0, 100mM NaCl, 5% glycerol). Purified full-length DDX41 protein, HBD, or HBD-WKK were added to the individual substrates in a serial dilution. Fluorescence polarization of the 6-FAM-labelled probes was analyzed on a Tecan Spark 20M plate reader at 20°C (excitation wavelength: 495 nm, emission wavelength: 520 nm, gain: 100, flashes: 15, integration time: 40 µs). Relative fluorescence polarization was calculated by subtracting the FP value of the oligo-only conditions. Binding constants (Kd values) were determined by fitting a Michaelis-Menten non-linear regression onto the relative FP values in GraphPad Prism (v7.04, Graphpad Software, Inc.).

4.2.12 Electrophoretic mobility shift assay

20 nM of 6-FAM-conjugated single- and double-stranded oligonucleotides were incubated with 25 µM of purified HBD or HBD-WKK mutant for 10 minutes at room temperature in interaction buffer (20 mM Tris-Cl pH 7.5, 150 mM NaCl, 10% glycerol, 1 mM EDTA, 1 mM DTT). 6× loading buffer (60% Glycerol, 20 mM Tris-Cl pH 8.0, 60 mM EDTA) was added to the samples before loading them on a 20% Novex TBE gel (Thermo Fisher Scientific). The gel was run for 45 minutes at 200 V in TBE buffer and scanned using a Typhoon FLA 9000 @ 473 nm to visualize the fluorescence of the 6-FAM-labelled probes.

4.2.13 FRET-based unwinding assay

RNA-DNA hybrid substrates with a single-stranded DNA overhang were generated by mixing an IBFQ-conjugated 38-mer DNA oligo (IDT) and a 6-FAM-conjugated 13-mer RNA oligo (IDT) and heating them to 95°C and gradually cooling them down to 4°C. Annealed substrates were incubated together with 5 μM ATP and either full-length DDX41 or mutant proteins. Increased fluorescence intensity upon addition of DDX41 after displacement of the quencher during unwinding was measured on a Spark M20 (Tecan) plate reader.

4.2.14 Cell viability assay

Cell viability assay was performed using the Cell Titer-Blue Cell Viability Assay (Promega) according to the manufacturer's instructions.

4.2.15 RNA-sequencing and data analysis

RNA was extracted using the RNeasy Plus Mini Kit (Qiagen). In brief, cells were lysed and genomic DNA was depleted. Samples were treated with DNase to remove residual DNA. After purification using spin columns, RNA was eluted in RNase-free water and stored at -80°C until library preparation. NGS library prep was performed with Illumina's TruSeq Stranded mRNA LT Sample Prep Kit following the TruSeq Stranded mRNA Reference Guide (Oct.2017) (Document # 1000000040498v00). Libraries were prepared with a starting amount of 1000 ng and amplified in 10 PCR cycles. Libraries were profiled in a High Sensitivity DNA on a 2100 Bioanalyzer (Agilent Technologies) and quantified using the Qubit dsDNA HS Assay Kit, in a Qubit 2.0 Fluorometer (Life Technologies). All 15 samples were pooled in equimolar ratio and sequenced on a NextSeq 500 High output FC, SR for 1×84 cycles plus 7 cycles for the index read. All genomic libraries were sequenced on an Illumina NextSeq 500 and de-multiplexed using blc2fastq (v2.19). RNA-seq samples were sequenced with a read length of 84 bp in single read mode. Samples were mapped using STAR (v2.7) against hg38 with the Gencode annotation (v25)[424,425]. Reads per gene were counted using featureCounts (v.1.6)[426]. The differential expression analysis was performed using Bioconductor (v2.46)/DESeq2 (v1.26)[427,428]. Genes were deemed significantly differentially regulated with an FDR below 1%. Coverage tracks were normalized and created using DeepTools (v3.4.1)[429]. Genes were deemed expressed within the analysis if they were tested for differential expression in the DESeq2 analysis. We used GSM2296622 to generate a list of expressed genes for the HCT116 cell line. The raw data were downloaded from GEO and mapped using STAR against hg38 with Gencode annotation. Reads per gene were counted using featureCounts.

4.2.16 MapR

MapR was performed according to the before published protocol with minor modifications[156,242]. Cells were either treated with indicated siRNAs or with 4 µM ActinomycinD D (Cell Signaling Technology) for 6 h. Concanavalin A-coated beads (Polysciences Europe GmbH) were activated in Binding Buffer (20 mM Hepes-KOH pH 7.9, 10 mM KCl, 1mM CaCl2, 1 mM MnCl2). 5*105 U2OS cells were washed twice with Wash Buffer (Hepes-NaOH pH7.5, 150 mM NaCl, 0.5 mM Spermidine, 1 mM protease inhibitor) at room temperature and afterward immobilized on the activated beads in 50 µl Wash Buffer containing 0.05% Digitonin. Either pA-MNase or RHΔ-MNase was added to the cells overnight at 4°C on a rotating wheel. After three washes with Wash Buffer containing 0.05% Digitonin (Millipore), samples resuspended in 100 µl Dig-Wash-Buffer were equilibrated on ice. The activity of the MNase was triggered by adding 2mM CaCl2 to the samples for 30 minutes. 2× Stop Buffer (68 µl 5M NaCl, 40 µl 0.5 M EDTA, 20 µl 0.2 M EGTA, 10 µl 5% digitonin, 5 µl 10mg/ml RNaseA) was mixed with the samples to stop the reaction. Chromatin fragments were released by incubating the samples for 20 minutes at 37°C and centrifugation at 16.000×g for 10 minutes at 4°C. Supernatants were incubated at 70°C in the presence of 0.1% SDS and 5 µg proteinase K. Before library preparation, the DNA was recovered by phenol-chloroform extraction. NGS library preparation was performed using NuGEN´s Ovation Ultralow System V2 (M01379 v5). Libraries were prepared with a starting amount of 1 ng of DNA and were amplified in 12 PCR cycles. Libraries were profiled in a High Sensitivity DNA on a 2100 Bioanalyzer (Agilent Technologies) and quantified using the Qubit dsDNA HS Assay Kit, in a Qubit 2.0 Fluorometer (Life technologies). All 18 samples were pooled in equimolar ratio and sequenced on one NextSeq 500 High output Flow cell, PE for 2x 42 cycles plus 8 cycles for the index read.

4.2.17 MapR analysis

U2OS, as well as HCT116 MapR samples, were mapped against h38 using bowtie2 (v2.3.4) the result was filtered for uniquely mapping reads[430]. Peak calling was performed using MACS2 (v2.1.2)[431] with the parameters "--keep-dup auto --broad --broad-cutoff 0.1 --bw 100 --min-length 100 --format BAMPE --g hs". The MapR samples were further analyzed using the FC between siDDX41 and siCtrl. A set of consensus regions was created using the intersection of peaks called per group (either in siDDX41 or siCtrl replicates). Then the union of these two peak sets was used to quantify the signal present in the samples. Using R/Bioconductor[428] packages the fold change for the consensus regions was calculated using the average normalized coverage (RPKM) of the

regions. Normalization was based on the total amount of sequenced reads. R-loop gains were determined based on the FC>2 in siDDX41 compared to siCtrl U2OS cells and FC>1.5 in siDDX41 compared to siCtrl HCT116 cells. Sequencing depth normalized coverage tracks for all samples and metagene/enrichment around the TSS or from TSS to TES were created using DeepTools (v3.4.1)[429] and further processed using custom R scripts.

4.2.18 sBLISS and data analysis

sBLISS in HCT116 cells was performed as previously described[354,355] with the following modifications: After blunting of DNA DSB ends in fixed nuclei, samples were 3' adenylated using Klenow Fragment (3'-->5' exo-) (NEB M0212) at a final amount of 60U per reaction in 1x NEBNext® dA-Tailing Reaction Buffer (NEB B6059). The A-tailing reaction was incubated at 37°C for 1 hour in a thermo-shaker at 300 rpm. Fixed nuclei were then washed 2× with CST buffer (CutSmart buffer B7204 supplemented with 0.1% Triton) to wash out the Klenow enzyme. DNA DSB-end labeling was performed as described[355] with the following modification: sBLISS linkers containing one thymine overhang at the 3' end of the reverse oligo were used. Downstream sample processing steps were carried out with 150 ng of DNA template input from each sample for in vitro transcription reaction. sBLISS data were processed as described previously[355] using GRCh38/hg38 reference genome with BWA-MEM98 (version 0.7.15). We used MACS2[431] (version 2.2.6) to call peaks from the BED files of UMI-DSB as reported previously[354]. Peaks identified by MACS2 with q-value < 0.01 were annotated using Chipseeker[432] (version 1.22.0). Peaks lists from both conditions were merged using bedtools[433] (version 2.27.0). The count-per-million (CPM) values for the merged peaks were calculated and normalized by library size with edgeR[434] (version 3.32.1). The peaks with gain and loss of breakage were classified based on the fold change greater than 2 between siCtrl and siDDX41 samples.

4.2.19 greenCUT&RUN and data analysis

CUT&RUN was performed in a stable U2OS cell line that expresses N-terminally GFP-tagged DDX41 under a doxycycline-inducible promoter. Expression of GFP-DDX41 was induced by adding 1 µg/ml doxycycline for 48 h or DMSO for un-induced control cells. Cells were mildly cross-linked with 1% formaldehyde for 2 min at room temperature. Quenching of the reaction with 125 mM glycine was followed by cell detachment using trypsin and two subsequent washes in Wash buffer (20 mM HEPES–KOH (pH 7.5), 150 mM NaCl and 0.5 mM spermidine and EDTA-free complete protease inhibitor). Concanavalin-A beads were activated in binding buffer (20 mM

HEPES–KOH (pH 7.9), 10 mM KCl, 1 mM CaCl2, and 1 mM MnCl2) for 5 min at room temperature, and afterward, 1*106 cells immobilized on the beads for 10 min at room temperature. After cell permeabilization with 0.05% digitonin-Wash buffer, 1 µg of GFP-nanobody-MNase (GFP nanobody LaG16 described in[435]) was added in 100 µl and incubated with the immobilized cells at 4°C for 30 min. Unbound MNase was washed out two times with digitonin-Wash buffer before transferring samples to an ice bath. The MNase was activated by the addition of 3 mM CaCl2 for 30 min and the reaction was subsequently stopped by adding 2× Stop buffer (340 mM NaCl, 20 mM EDTA, 10 mM EGTA, 0.05% digitonin, 100 µg/ml of RNase A and 50 mg/ml glycogen). DNA fragments were released for 20 min at 37°C before de-crosslinking overnight at 55°C in the presence of 0.1% SDS and 1.5 µl of 20mg/ml proteinase K. DNA was cleaned up by phenol-chloroform extraction before subsequent library preparation using the Accel-NGS 1S Plus DNA Library Kit (Swift Bioscience) according to the manufactures' protocol suggested for the retention of small fragments (>40 bp). Libraries were dual indexed and amplified for 14 cycles (NEBNext Multiplex Oligos for Illumina, Dual Index Primers Set 1). An equimolar pool of libraries was prepared and further purified away from primer and adaptor dimers on a 2% agarose gel using the Zymoclean Gel DNA Recovery Kit (Zymo). Final quantification and quality control before sequencing was done on an Agilent 2200 Tapestation System. Samples were sequenced in paired read mode with 34 bp in read 1 and 49 bp in read 2. The first 15 bases of the second read of the cut and run data were removed and the data was adapter trimmed using Cutadapt (v.1.18)[436]. The data was mapped against hg38 using bowtie2 (v.2.3.4)[430] and filtered for uniquely mapping reads. Peak calling was done using MACS2 (v2.1.2)[431] with the following parameters "-g hs --min-length 150 --format BAMPE --keep-dup auto".

4.2.20 DNA fiber spreading

U2OS cells were labeled with 5-Chloro-2′-deoxyuridine (CldU, 30 µM) for 30 min, washed once with warm PBS, then labeled for 30 min with 5-Iodo-2′-deoxyuridine (IdU, 340 µM). Cells were either transfected with siDDX41/siCtrl for 24 h or treated with 0.1 µM aphidicolin (APH) for 1.5 h. After labeling, cells were washed once with warm and 3× with cold PBS, then trypsinized and spun down (300×g, 5 min). They were resuspended in cold PBS, counted, and diluted to 5×105/ml. Labeled cells were diluted with twice the number of unlabeled cells. 4 µl of the cell suspension were mixed with 7.5 µl of the lysis buffer (200mM Tris-HCl pH 7.4, 50 mM EDTA, 0.5% SDS) directly on the SuperFrost Plus microscopy slide (Thermo Scientific) and incubated horizontally for 9 min. The slides were then tilted at 30°−45°, allowing DNA fibers to spread to the bottom of

the slide. DNA spreads were air-dried and fixed with 3:1 methanol:acetic acid overnight at 4°C. The fibers were rehydrated 3×3 min in PBS, dipped once in Milli-Q water and denatured in 2.5 M HCl for 1.5 h at RT, then washed 5×3 min in PBS. The slides were blocked for 40 min in the blocking solution (2% BSA, 0.1% Tween 20 in PBS) and incubated with primary antibodies (mouse anti-BrdU, 1:100, BD Bioscience and rat anti-BrdU, 1:500, Abcam) at RT for 2.5 h. After 3×5 min washes with PBS-T, the slides were incubated with secondary antibodies (goat anti-mouse Cy3.5, Abcam and goat anti-rat Cy5, Abcam) at RT in the dark for 1 h. The spreads were washed 3×5 min with PBS-T, dipped in Milli-Q water, and air-dried completely in the dark. The slides were mounted using Prolong Gold AntiFade mountant (Thermo Scientific), imaged with Visiscope 5-Elements Spinning Disc Confocal microscope (Visitron Systems, Germany) (magnification: 60× Water immersion objective with 2x extra magnification; laser lines and corresponding emission filters: 640nm, 692/40 and 561nm 623/32) and quantified using the Fiji/ImageJ software[423].

4.2.21 Immunofluorescence

Cells were washed 2× with PBS, incubated with 0.4% NP-40 for 20 or 40 min on ice, and washed 2× with PBS-T (0.1%). Cells were fixed with 4% paraformaldehyde in PBS for 15 min at room temperature, washed 2× with PBS-T (0.1%), and permeabilized with Triton-X-100 (0.3%) for 5 min at room temperature, followed by 2× washes with PBS. Cells were blocked for 1 hour with 5% fetal bovine serum albumin in PBS-T (0.1%) containing penicillin and streptomycin. Incubation with primary antibodies diluted in blocking buffer was performed overnight at 4°C and followed by 3× washes with PBS-T and 1-hour incubation with Alexa Fluor-coupled secondary antibodies in a dark chamber at room temperature. Nuclei were counterstained with 1µg/ml Hoechst33342 in PBS either simultaneously with secondary antibody incubation or for 30 minutes. For chromatin retention assay permeabilization, blocking and antibody incubations steps were omitted. Cells were washed 2× with PBS-T and kept at 4°C in PBS until imaging. Imaging was performed with an Opera Phenix (PerkinElmer) microscope using a 40× 1.1NA water objective. Image analysis was performed by using Harmony High-Content Imaging and Analysis Software (version 4.4, PerkinElmer). Standard building blocks allowed for nuclei segmentation based on the Hoechst signal and cells on the edges of the field were excluded. Mean intensity measurements were performed for maximum projections and spot detection was calculated by using algorithm B.

4.2.22 Hematopoietic stem and progenitor cell experiments

For DDX41 knockdown experiments primary human CD34+ cells (from cord blood, purchased from Lonza) were transfected with two different shRNA constructs (TL305064C; GCTATGCAGACCAAGCAGGTCAGCAACAT; TL305064D; GCGTGCGGAAGAAATACCACATCCTGGTG) (Origene) using Amaxa® Human CD34+ Cell Nucleofector® Kit according to the manufacturer's recommendations (Lonza). Similarly, for ectopic expression of DDX41 WT, DDX41 L237F P238T, and DDX41 R525H, respectively, primary human CD34+ cells (from cord blood, purchased from Lonza) were transfected using Amaxa® Human CD34+ Cell Nucleofector® Kit according to the manufacturer's recommendations (Lonza). Cells were cultured in StemSpan SFEM II supplemented with myeloid expansion supplement containing SCF, TPO, G-CSF, and GM-CSF (Stemcell Technologies) for 24 h before isolation of GFP+ cells with a BD FACSMelody cell sorter, using double sorting to ensure maximum purity (BD Biosciences). For immunofluorescence experiments, cells were seeded onto glass slides, fixed with 4% PFA for 10 min, permeabilized using 0.15% Triton X-100 for 2 min, and blocked in 1% BSA/PBS. 53BP1 was detected using anti-53BP1 (nb100-904; Novus Biologicals) followed by Alexa Fluor® 594-conjugated goat anti-rabbit antibody (Thermo Fisher Scientific). Slides were mounted in VectaShield containing 1 µg/ml DAPI (Vector Laboratories). Images were acquired on a DMi8 Leica inverted microscope (100x objective) and processed using LasX software (Leica). Quantification of mean fluorescence intensity (MFI) was performed using the following formula: mean fluorescence of selected cell – (area of selected cell × mean fluorescence of background readings). Values are displayed as arbitrary units (A.U.). To assess DDX41 knockdown efficiency, about 5000 eGFP+ cells were sorted in RLT buffer (Qiagen) for RNA extraction followed by cDNA synthesis. DDX41 expression was quantified using qRT-PCR (DDX41-fw-qPCR, DDX41-rev-qPCR).

4.2.23 Protein production

His6-3C-DDX41 full-length WT and R525H were expressed in SF9 insect cells using the Bac-to-Bac system and SF900 III media (Thermo Fisher). His6-DDX41 (153-410), His6-GST-3C-RNaseH (27-76 = HBD)-AVI-tag WT, W43/K59/K60-A (WKK-A) (all pET28), and pelB-GFP nanobody (LaG16)-MNase-His6-HA (pET21b, LaG16 nanobody described in X) were expressed in E.coli (BL21 DE3 codon+) using LB media. Cells were lysed in lysis buffer (30 mM Tris-Cl pH 8.0, 500 mM NaCl, 1 mM MgCl2, 1 mM DTT, 5% glycerol, 100 U/ml benzonase, EDTA-free cOmplete protease inhibitor cocktail, 15 mM imidazole except for the HBD constructs, 0.5% Triton

X-100 for the DDX41 full-length constructs), using a Branson Sonifier 450 and cleared by centrifugation (40000×g, 30 min at 4°C). In the case of the HBD constructs, an additional 500 mM NaCl was added to the cleared lysates and PEI-based precipitation of nucleic acids (0.2% w/v Polyethylenimine, 40 kDa, pH 7.4) for 5 min at 4°C was performed, followed by a second round of centrifugation (4000 x g, 4°C, 15 min). Recombinant proteins were affinity-purified from cleared lysates using an NGC Quest Plus FPLC system (BioRad) and Cytiva columns: HisTrap FF crude (DDX41 FL variants), HisTrap FF 5 ml (DDX41 (153-410), LaG16-MNase), GSTrap HP 5 ml (HBD variants), following the manufacturer's protocols. DDX41(153-410) and LaG16-MNase were further subjected to Heparin-based chromatography (HisTrap Heparin HP 5 ml, Cytiva, in 30 mM Na-Hepes, 25 mM NaCl, 5% glycerol) following the manufacturer's protocol. DDX41 FL variants and HBDs were digested with His6-3C protease (1:100 w/w) overnight at 4°C in the presence of 1 mM DTT to cleave off the His6- and His6-GST tag, respectively. Digested HBDs were run over a HisTrap ff 5 ml column (Cytiva) to absorb out the His6-GST and His6-3C protease. All recombinant proteins were concentrated using Amicon spin concentrators (Merck Millipore) and subjected to gel filtration (in 30 mM Na-Hepes, 300 mM NaCl, 1 mM DTT, 10% Glycerol, pH 7.5, additional 1 mM EDTA for HBDs and Lag16-MNase). DDX41 FL variants were run twice on a Superdex 200 16/60 pg (Cytiva), and all other proteins were run once on a Superdex 75 16/60 pg (Cytiva). Peak fractions containing the recombinant proteins after gel filtration were pooled and protein concentration was determined by using absorbance spectroscopy and the respective extinction coefficient at 280 nm before aliquots were flash-frozen in liquid nitrogen and stored at -80°C.

4.2.24 Fluorescence-activated cell sorting

Expression of N-terminally GFP-tagged DDX41 WT, L237F+P238T, or R525H in OCI-AML3 cells was induced with 3 µg/ml doxycycline (Sigma-Aldrich). 72 h after induction, cells were spun down and washed twice with PBS. After re-suspending cells in PBS, they were sorted by FACS using a 100 µM nozzle on a BD FACSAria III SORP (Becton Dickinson) in purity precision mode with FACSDiva software version 8.0.2. Cells of interest were identified via FSC-A/SSC-A. Subsequently, doublets were excluded via FSC-A/FSC-H and dead cells were excluded by DAPI staining (0.5 µg/ml final concentration) using a 405 nm laser and 450/50 BP. GFP cutoff was set according to non-expressing cells. Sorting based on GFP was achieved using the 488 nm laser and 530/30 band pass filter. Roughly 500.000 cells were sorted directly into fresh a-MEM containing 20% FBS and further cultured until subsequent experiments.

5 Abbreviations

Abbreviation	Full form
ADP	Adenosine diphosphate
ACN	Acetonitrile
AID	Activation-induced cytidine deaminase
ALT	Alternative lengthening of telomeres
ALS4	Amyotrophic lateral sclerosis 4
AML	Acute myeloid leukemia
AOA2	Ataxia with oculomotor apraxia type 2
APH	Aphidicolin
ATP	Adenosine triphosphate
AURKA	Aurora kinase A
AURKB	Aurora kinase B
BER	Base excision repair
bp	Base pair
BP	Biological process
BSA	Bovine serum albumin
CD	Co-directional
CHK1	Checkpoint kinase 1
cGAS	Cyclic GMP-AMP synthase
CGI	CpG island promoter
CPC	Chromosome passenger complex
CPT	Camptothecin
CRISPR	Clustered regularly interspaced short palindromic repeats
CSR	Class-switch recombination
D-loop	Displacement loop
dilncRNA	Damage-induced long-non-coding RNA
DMSO	Dimethyl sulfoxide
DNMT	DNA methyltransferases
DOX	Doxycycline
DSB	Double strand break

dsDNA	Double-stranded DNA
dsRNA	Double-stranded RNA
EDTA	Ethylenediaminetetraacetic acid
EMSA	Electron mobility shift assay
EU	5-Ethynyl uridine
FC	Fold change
FDR	False discovery rate
FP	Fluorescence polarization assay
G4	G-quadruplex
GADD45A	Growth arrest and DNA damage protein 45A
gRNA	Guide RNA
HBD	Hybrid binding domain
HO	Head-on
HP1y	Hetero-chromatin protein 1y
HR	Homologous recombination
HSPC	Hematopoietic stem and progenitor cell
IFN	Interferon
IRF3	Interferon regulatory factor 3
kb	Kilo base
LC	Liquid chromatography
lncRNA	Long non-coding RNA
m^6A	N^6-methyladenosine
MCM	Minichromosome maintenance complex
Mb	Mega bases
MS	Mass spectrometry
mtDNA	Mitochondrial DNA
NELF	Negative elongation factor
NHEJ	Non-homologous end-joining
OGRE	Origin G-rich repeated element
ORC	Origin recognition complex
PBS	Phosphate buffered saline
PCNA	Proliferating cell nuclear antigen
PFA	Paraformaldehyde
PIC	Pre-initiation complex

PladB	Pladienolide B
PRC2	Polycomb repressive complex
pTEFb	Positive transcription elongation factor-b
rDNA	Ribosomal DNA
RPA	Replication protein A
RFC	Replication factor C
RNA Pol	RNA polymerase
RNP	Ribonucleoprotein
ROS	Reactive oxygen species
RUNX1	Runt-related transcription 1
sBLISS	Break Labeling In Situ and Sequencing
SETX	Senataxine
SILAC	Stable isotope labeling in cell culture
ssDNA	Single-stranded DNA
STING	Stimulator of interferon genes
TC-NER	Transcription-coupled nucleotide excision repair
TFA	Trifluoroacetic acid
TonEBP	Tonicity-responsive enhancer binding protein
TRC	Transcription-replication conflict
tRNA	Transfer-RNA
TSS	Transcription start site
TTS	Transcription termination site

6 References

1. Masai, H., Matsumoto, S., You, Z., Yoshizawa-Sugata, N. & Oda, M. Eukaryotic chromosome DNA replication: Where, when, and how? *Annu. Rev. Biochem.* **79**, 89–130 (2010).

2. Zeman, M. K. & Cimprich, K. A. Causes and consequences of replication stress. *Nat. Cell Biol.* **16**, 2–9 (2014).

3. Vashee, S. *et al.* Sequence-independent DNA binding and replication initiation by the human origin recognition complex. *Genes Dev.* **17**, 1894–1908 (2003).

4. Vashee, S., Simancek, P., Challberg, M. D. & Kelly, T. J. Assembly of the human origin

recognition complex. *J. Biol. Chem.* **276**, 26666–26673 (2001).

5. Cheng, J. *et al.* Structural insight into the assembly and conformational activation of human origin recognition complex. *Cell Discov.* **6**, (2020).

6. Remus, D. *et al.* Concerted loading of Mcm2-7 double hexamers around DNA during DNA replication origin licensing. *Cell* **139**, 719–730 (2009).

7. Evrin, C. *et al.* A double-hexameric MCM2-7 complex is loaded onto origin DNA during licensing of eukaryotic DNA replication. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 20240–20245 (2009).

8. Moyer, S. E., Lewis, P. W. & Botchan, M. R. Isolation of the Cdc45/Mcm2-7/GINS (CMG) complex, a candidate for the eukaryotic DNA replication fork helicase. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 10236–10241 (2006).

9. Blow, J. J., Ge, X. Q. & Jackson, D. A. How dormant origins promote complete genome replication. *Trends Biochem. Sci.* **36**, 405–414 (2011).

10. Muzi-Falconi, M., Giannattasio, M., Foiani, M. & Plevani, P. The DNA polymerase alpha-primase complex: multiple functions and interactions. *ScientificWorldJournal.* **3**, 21–33 (2003).

11. Pellegrini, L. The Pol α-primase complex. *Subcell. Biochem.* **62**, 157–169 (2012).

12. Daigaku, Y. *et al.* A global profile of replicative polymerase usage. *Nat. Struct. Mol. Biol.* **22**, 192–198 (2015).

13. Yeeles, J. T. P., Janska, A., Early, A. & Diffley, J. F. X. How the Eukaryotic Replisome Achieves Rapid and Efficient DNA Replication. *Mol. Cell* **65**, 105–116 (2017).

14. Yu, C. *et al.* Strand-Specific Analysis Shows Protein Binding at Replication Forks and PCNA Unloading from Lagging Strands when Forks Stall. *Mol. Cell* **56**, 551–563 (2014).

15. Nick McElhinny, S. A., Gordenin, D. A., Stith, C. M., Burgers, P. M. J. & Kunkel, T. A. Division of Labor at the Eukaryotic Replication Fork. *Mol. Cell* **30**, 137–144 (2008).

16. Burgers, P. M. J. Polymerase dynamics at the eukaryotic DNA replication fork. *J. Biol. Chem.* **284**, 4041–4045 (2009).

17. Boehm, E. M., Gildenberg, M. S. & Washington, M. T. The Many Roles of PCNA in Eukaryotic DNA Replication. *Enzym.* **39**, 231–254 (2016).

18. Kelman, Z. PCNA: structure, functions and interactions. *Oncogene* **14**, 629-640. (1997).

19. Kang, M. S. *et al.* Regulation of PCNA cycling on replicating DNA by RFC and RFC-like complexes. *Nat. Commun.* **10**, (2019).

20. Zou, Y., Liu, Y., Wu, X. & Shell, S. M. Functions of human replication protein A (RPA): from DNA replication to DNA damage and stress responses. *J. Cell. Physiol.* **208**, 267–273 (2006).

21. Acevedo, J., Yan, S. & Michae, W. M. Direct binding to replication protein A (RPA)-coated single-stranded DNA allows recruitment of the ATR activator topBP1 to sites of DNA damage. *J. Biol. Chem.* **291**, 13124–13131 (2016).

22. Lee, Z. & J., E. S. Sensing DNA Damage Through ATRIP Recognition of RPA-ssDNA Complexes. *Science (80-.).* **300**, 1542–1548 (2003).

23. Nam, E. A. & Cortez, D. ATR signalling: more than meeting at the fork. *Biochem. J.* **436**, 527–536 (2011).

24. Maréchal, A. & Zou, L. DNA damage sensing by the ATM and ATR kinases. *Cold Spring Harb. Perspect. Biol.* **5**, 1–18 (2013).

25. Koundrioukoff, S. *et al.* Stepwise Activation of the ATR Signaling Pathway upon Increasing Replication Stress Impacts Fragile Site Integrity. *PLoS Genet.* **9**, (2013).

26. Petermann, E. & Helleday, T. Pathways of mammalian replication fork restart. *Nat. Rev. Mol. Cell Biol.* **11**, 683–687 (2010).

27. Quinet, A., Lemaçon, D. & Vindigni, A. Replication Fork Reversal: Players and Guardians. *Mol. Cell* **68**, 830–833 (2017).

28. Woodward, A. M. *et al.* Excess Mcm2-7 license dormant origins of replication that can be used under conditions of replicative stress. *J. Cell Biol.* **173**, 673–683 (2006).

29. Elvers, I., Johansson, F., Groth, P., Erixon, K. & Helleday, T. UV stalled replication forks restart by re-priming in human fibroblasts. *Nucleic Acids Res.* **39**, 7049–7057 (2011).

30. Powers, K. T. & Washington, M. T. Eukaryotic translesion synthesis: Choosing the right tool for the job. *DNA Repair (Amst).* **71**, 127–134 (2018).

31. Branzei, D. & Foiani, M. Template Switching: From Replication Fork Repair to Genome Rearrangements. *Cell* **131**, 1228–1230 (2007).

32. Ying, S., Hamdy, F. C. & Helleday, T. Mre11-dependent degradation of stalled DNA replication forks is prevented by BRCA2 and PARP1. *Cancer Res.* **72**, 2814–2821 (2012).

33. Sainsbury, S., Bernecky, C. & Cramer, P. Structural basis of transcription initiation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* **16**, 129–143 (2015).

34. Paule, M. R. & White, R. J. Survey and summary: transcription by RNA polymerases I and III. *Nucleic Acids Res.* **28**, 1283–1298 (2000).

35. Hoffmann, N. A., Jakobi, A. J., Vorländer, M. K., Sachse, C. & Müller, C. W. Transcribing RNA polymerase III observed by electron cryomicroscopy. *FEBS J.* **283**, 2811–2819 (2016).

36. Neyer, S. *et al.* Structure of RNA polymerase I transcribing ribosomal DNA genes. *Nature* **540**, 607–610 (2016).

37. Meyer, P. A., Ye, P., Suh, M.-H., Zhang, M. & Fu, J. Structure of the 12-subunit RNA polymerase II refined with the aid of anomalous diffraction data. *J. Biol. Chem.* **284**, 12933–12939 (2009).

38. Bushnell, D. A. & Kornberg, R. D. Complete, 12-subunit RNA polymerase II at 4.1-Å resolution: Implications for the initiation of transcription. *Proc. Natl. Acad. Sci.* **100**, 6969 LP – 6973 (2003).

39. Schier, A. C. & Taatjes, D. J. Structure and mechanism of the RNA polymerase II transcription machinery. *Genes Dev.* **34**, 465–488 (2020).

40. Buratowski, S., Hahn, S., Guarente, L. & Sharp, P. A. Five intermediate complexes in transcription initiation by RNA polymerase II. *Cell* **56**, 549–561 (1989).

41. Dienemann, C., Schwalb, B., Schilbach, S. & Cramer, P. Promoter Distortion and Opening in the RNA Polymerase II Cleft. *Mol. Cell* **73**, 97-106.e4 (2019).

42. Schilbach, S. *et al.* Structures of transcription pre-initiation complex with TFIIH and Mediator. *Nature* **551**, 204–209 (2017).

43. Yan, C. *et al.* Transcription preinitiation complex structure and dynamics provide insight into genetic diseases. *Nat. Struct. Mol. Biol.* **26**, 397–406 (2019).

44. Grünberg, S. & Hahn, S. Structural insights into transcription initiation by RNA polymerase II. *Trends Biochem. Sci.* **38**, 603–611 (2013).

45. Vázquez, M. E., Caamaño, A. M. & Mascareñas, J. L. From transcription factors to designed sequence-specific DNA-binding peptides. *Chem. Soc. Rev.* **32**, 338–349 (2003).

46. Martínez-Balbás, M. A., Dey, A., Rabindran, S. K., Ozato, K. & Wu, C. Displacement of sequence-specific transcription factors from mitotic chromatin. *Cell* **83**, 29–38 (1995).

47. Pan, S., Czarnecka-Verner, E. & Gurley, W. B. Role of the TATA binding protein-transcription

factor IIB interaction in supporting basal and activated transcription in plant cells. *Plant Cell* **12**, 125–136 (2000).

48. Allen, B. L. & Taatjes, D. J. The Mediator complex: A central integrator of transcription. *Nat. Rev. Mol. Cell Biol.* **16**, 155–166 (2015).

49. Plank, J. L. & Dean, A. Enhancer Function: Mechanistic and Genome-Wide Insights Come Together. *Mol. Cell* **55**, 5–14 (2014).

50. Kagey, M. H. *et al.* Mediator and cohesin connect gene expression and chromatin architecture. *Nature* **467**, 430–435 (2010).

51. Eick, D. & Geyer, M. The RNA Polymerase II Carboxy-Terminal Domain (CTD) Code. (2013).

52. Adelman, K. & Lis, J. T. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat. Rev. Genet.* **13**, 720–731 (2012).

53. Core, L. J. & Lis, J. T. Transcription regulation through promoter-proximal pausing of RNA polymerase II. *Science* **319**, 1791–1792 (2008).

54. Jonkers, I. & Lis, J. T. Getting up to speed with transcription elongation by RNA polymerase II. *Nat. Rev. Mol. Cell Biol.* **16**, 167–177 (2015).

55. Yamaguchi, Y., Inukai, N., Narita, T., Wada, T. & Handa, H. Evidence that negative elongation factor represses transcription elongation through binding to a DRB sensitivity-inducing factor/RNA polymerase II complex and RNA. *Mol. Cell. Biol.* **22**, 2918–2927 (2002).

56. Peterlin, B. M. & Price, D. H. Controlling the elongation phase of transcription with P-TEFb. *Mol. Cell* **23**, 297–305 (2006).

57. Zhou, Q., Li, T. & Price, D. H. RNA polymerase II elongation control. *Annu. Rev. Biochem.* **81**, 119–143 (2012).

58. Laitem, C. *et al.* CDK9 inhibitors define elongation checkpoints at both ends of RNA polymerase II-transcribed genes. *Nat. Struct. Mol. Biol.* **22**, 396–403 (2015).

59. Bacon, C. W. & D'Orso, I. CDK9: a signaling hub for transcriptional control. *Transcription* **10**, 57–75 (2019).

60. Hsin, J.-P. & Manley, J. L. The RNA polymerase II CTD coordinates transcription and RNA processing. *Genes Dev.* **26**, 2119–2137 (2012).

61. Price, D. H. Transient pausing by RNA polymerase II. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 4810–

4812 (2018).

62. Buratowski, S. Progression through the RNA Polymerase II CTD Cycle. *Mol. Cell* **36**, 541–546 (2009).

63. Bowman, E. A. & Kelly, W. G. RNA polymerase II transcription elongation and Pol II CTD Ser2 phosphorylation: A tail of two kinases. *Nucleus* **5**, 224–236 (2014).

64. Veloso, A. *et al.* Rate of elongation by RNA polymerase II is associated with specific gene features and epigenetic modifications. *Genome Res.* **24**, 896–905 (2014).

65. Sims, R. J., Belotserkovskaya, R. & Reinberg, D. Elongation by RNA polymerase II: The short and long of it. *Genes Dev.* **18**, 2437–2468 (2004).

66. Fong, N. *et al.* Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. *Genes Dev.* **28**, 2663–2676 (2014).

67. Aslanzadeh, V., Huang, Y., Sanguinetti, G. & Beggs, J. D. Transcription rate strongly affects splicing fidelity and cotranscriptionality in budding yeast. *Genome Res.* **28**, 203–213 (2018).

68. Saldi, T., Riemondy, K., Erickson, B. & Bentley, D. L. Alternative RNA structures formed during transcription depend on elongation rate and modify RNA processing. *Mol. Cell* **81**, 1789-1801.e5 (2021).

69. Castillo-guzman, D., Hartono, S. R., Sanz, L. A. & Chédin, F. SF3B1-targeted Splicing Inhibition Triggers Global Alterations in Transcriptional Dynamics and R-Loop Metabolism. (2020).

70. Porrua, O. & Libri, D. Transcription termination and the control of the transcriptome: Why, where and how to stop. *Nat. Rev. Mol. Cell Biol.* **16**, 190–202 (2015).

71. Reyes, A. & Huber, W. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res.* **46**, 582–592 (2018).

72. Castillo-Guzman, D. & Chédin, F. Defining R-loop classes and their contributions to genome instability. *DNA Repair (Amst).* **106**, 103182 (2021).

73. Niehrs, C. & Luke, B. Regulatory R-loops as effectors of gene expression and genome stability. *Nat Rev Mol Cell Biol.* **21**, 167–178 (2021).

74. Crossley, M. P., Bocek, M. & Cimprich, K. A. R-Loops as Cellular Regulators and Genomic Threats. *Mol. Cell* **73**, 398–411 (2019).

75. Kasamatsu, H., Robberson, D. L. & Vinograd, J. A Novel Closed-Circular Mitochondrial DNA

with Properties of a Replicating Intermediate. **68**, 2252–2257 (1971).

76. Robberson, D. L., Kasamatsu, H. & Vinograd, J. Replication of mitochondrial DNA. Circular replicative intermediates in mouse L cells. *Proc. Natl. Acad. Sci. U. S. A.* **69**, 737–741 (1972).

77. Thomas, M., White, R. L. & Davis, R. W. Hybridization of RNA to double stranded DNA: Formation of R loops. *Proc. Natl. Acad. Sci. U. S. A.* **73**, 2294–2298 (1976).

78. Berget, S. M., Moore, C. & Sharp, P. A. Spliced segments at the 5 ' terminus of adenovirus 2 late mRNA *. **74**, 3171–3175 (1977).

79. Chow, L. T., Gelinas, R. E. & Roberts, R. J. An Amazing Sequence Arrangement of Adenovirus RNA at the 5 ' Ends. **12**, 1–8 (1977).

80. Boguslawski, S. J. *et al.* Characterization of monoclonal antibody to D N A • R N A and its application to immunodetection of hybrids. **89**, 123–130 (1986).

81. Drolet, M. *et al.* Overexpression of RNase H partially complements the growth defect of an Escherichia coli AtopA mutant : R-loop formation is a major problem in the absence of DNA topoisomerase I. **92**, 3526–3530 (1995).

82. Roberts, R. W. & Crothers, D. M. Stability and Properties of Double and Triple Helices : Dramatic Effects of RNA or DNA Backbone Composition. **258**, (1992).

83. Shaw, N. N. & Arya, D. P. Recognition of the unique structure of DNA:RNA hybrids. *Biochimie* **90**, 1026–1039 (2008).

84. Lesnik, E. A. & Freier, S. M. Relative Thermodynamic Stability of DNA , RNA , and DNA : RNA Hybrid Duplexes : Relationship with Base Composition and Structure. 10807–10815 (1995).

85. Gyi, J. I., Lane, A. N., Conn, G. L. & Brown, T. The orientation and dynamics of the C2 ′ -OH and hydration of RNA and DNA · RNA hybrids. **26**, 3104–3110 (1998).

86. Duquette, M. L., Handa, P., Vincent, J. A., Taylor, A. F. & Maizels, N. Intracellular transcription of G-rich DNAs induces formation of G-loops , novel structures containing G4 DNA. 1618–1629 (2004) doi:10.1101/gad.1200804.led.

87. Malig, M., Hartono, S. R., Giafaglione, J. M., Sanz, L. A. & Chedin, F. High-Throughput Single-Molecule R-loop Footprinting Reveals Principles of R-loop Formation. *bioRxiv* (2019) doi:10.1101/640094.

88. Skourti-stathaki, K. & Proudfoot, N. J. A double-edged sword : R loops as threats to genome

integrity and powerful regulators of gene expression. 1384–1396 (2014) doi:10.1101/gad.242990.114.Freely.

89. Sanz, L. A. *et al.* Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals. *Mol. Cell* **63**, 167–178 (2016).

90. De Magis, A. *et al.* DNA damage and genome instability by G-quadruplex ligands are mediated by R loops in human cancer cells. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 816–825 (2019).

91. Miglietta, G., Russo, M. & Capranico, G. G-quadruplex – R-loop interactions and the mechanism of anticancer G-quadruplex binders. 1–16 (2020) doi:10.1093/nar/gkaa944.

92. Graf, M. *et al.* Telomere Length Determines TERRA and R-Loop Regulation through the Cell Cycle. *Cell* **170**, 72-85.e14 (2017).

93. El Hage, A., Webb, S., Kerr, A. & Tollervey, D. Genome-Wide Distribution of RNA-DNA Hybrids Identifies RNase H Targets in tRNA Genes, Retrotransposons and Mitochondria. *PLoS Genet.* **10**, (2014).

94. Balk, B. *et al.* Telomeric RNA-DNA hybrids affect telomere-length dynamics and senescence. *Nat. Struct. &Amp; Mol. Biol.* **20**, 1199 (2013).

95. Kabeche, L., Nguyen, H. D., Buisson, R. & Zou, L. A mitosis-specific and R loop-driven ATR pathway promotes faithful chromosome segregation. *Science (80-.).* (2018) doi:10.1126/science.aan6490.

96. Chen, J. Y., Zhang, X., Fu, X. D. & Chen, L. R-ChIP for genome-wide mapping of R-loops by using catalytically inactive RNASEH1. *Nat. Protoc.* **14**, 1661–1685 (2019).

97. Roy, D., Zhang, Z., Lu, Z., Hsieh, C.-L. & Lieber, M. R. Competition between the RNA Transcript and the Nontemplate DNA Strand during R-Loop Formation In Vitro: a Nick Can Serve as a Strong R-Loop Initiation Site. *Mol. Cell. Biol.* **30**, 146–159 (2010).

98. Rinaldi, C., Pizzul, P., Longhese, M. P. & Bonetti, D. Sensing R-Loop-Associated DNA Damage to Safeguard Genome Stability. *Front. Cell Dev. Biol.* **8**, 1–13 (2021).

99. Bader, A. S. & Bushell, M. DNA:RNA hybrids form at DNA double-strand breaks in transcriptionally active loci. *Cell Death Dis.* **11**, (2020).

100. Ngo, G. H. P., Grimstead, J. W. & Baird, D. M. UPF1 promotes the formation of R loops to stimulate DNA double-strand break repair. *Nat. Commun.* **12**, 1–15 (2021).

101. Cohen, S. *et al.* Senataxin resolves RNA:DNA hybrids forming at DNA double-strand breaks to prevent translocations. *Nat. Commun.* **9**, (2018).

102. Aguilera, A. & García-Muse, T. R Loops: From Transcription Byproducts to Threats to Genome Stability. *Mol. Cell* **46**, 115–124 (2012).

103. Allison, D. F. & Wang, G. G. R-loops: Formation, function, and relevance to cell stress. *Cell Stress* **3**, 38–46 (2019).

104. Hage, A. El, French, S. L., Beyer, A. L. & Tollervey, D. Loss of Topoisomerase I leads to R-loop-mediated transcriptional blocks during ribosomal RNA synthesis. 1546–1558 (2010) doi:10.1101/gad.573310.Freely.

105. Chan, Y. A. *et al.* Genome-Wide Profiling of Yeast DNA:RNA Hybrid Prone Sites with DRIP-Chip. *PLoS Genet.* **10**, (2014).

106. Rivosecchi, J. *et al.* Senataxin homologue Sen1 is required for efficient termination of RNA polymerase III transcription. *EMBO J.* **38**, 1–14 (2019).

107. Chédin, F. Nascent Connections: R-Loops and Chromatin Patterning. *Trends Genet.* **32**, 828–838 (2016).

108. Jiang, F. *et al.* Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. **351**, (2016).

109. Szczelkun, M. D. *et al.* Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 9798–9803 (2014).

110. Wahba, L., Gore, S. K. & Koshland, D. The homologous recombination machinery modulates the formation of RNA-DNA hybrids and associated chromosome instability. *Elife* **2013**, 1–20 (2013).

111. Cloutier, S. C. *et al.* Regulated Formation of lncRNA-DNA Hybrids Enables Faster Transcriptional Induction and Environmental Adaptation. *Mol. Cell* **61**, 393–404 (2016).

112. Moran, E. C. *et al.* Mitotic R-loops direct Aurora B kinase to maintain centromeric cohesion. *bioRxiv* 2021.01.14.426738 (2021).

113. Chen, C., Pohl, T. J., Chan, A., Slocum, J. S. & Zakian, V. A. Saccharomyces cerevisiae Centromere RNA Is. **213**, 465–479 (2019).

114. Yu, K., Chedin, F., Hsieh, C. L., Wilson, T. E. & Lieber, M. R. R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells. *Nat. Immunol.* **4**, 442–451 (2003).

115. Lieber, M. R. Site-specific recombination in the immune system 1 . *FASEB J.* **5**, 2934–2944 (1991).

116. Matthews, A. J., Zheng, S., DiMenna, L. J. & Chaudhuri, J. Regulation of immunoglobulin class-switch recombination: Choreography of noncoding transcription, targeted DNA deamination, and long-range DNA repair. in *Advances in Immunology* vol. 122 1–57 (Academic Press Inc., 2014).

117. Stavnezer, J. *et al.* Immunoglobulin heavy-chain switching may be directed by prior induction of transcripts from constant-region genes. *Proc. Natl. Acad. Sci. U. S. A.* **85**, 7704–7708 (1988).

118. Reaban, M. E. & Griffin, J. A. Induction of RNA-stabilized DMA conformers by transcription of an immunoglobulin switch region. *Nature* **348**, 342–344 (1990).

119. Ribeiro de Almeida, C. *et al.* RNA Helicase DDX1 Converts RNA G-Quadruplex Structures into R-Loops to Promote IgH Class Switch Recombination. *Mol. Cell* **70**, 650-662.e8 (2018).

120. Chaudhuri, J. & Alt, F. W. Class-switch recombination: Interplay of transcription, DNA deamination and DNA repair. *Nat. Rev. Immunol.* **4**, 541–552 (2004).

121. Santos-Pereira, J. M. & Aguilera, A. R loops: New modulators of genome dynamics and function. *Nat. Rev. Genet.* **16**, 583–597 (2015).

122. Muramatsu, M. *et al.* Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell* **102**, 553–563 (2000).

123. Revy, P. *et al.* Activation-induced cytidine deaminase (AID) deficiency causes the autosomal recessive form of the hyper-IgM syndrome (HIGM2). *Cell* **102**, 565–575 (2000).

124. Zheng, S. *et al.* Non-coding RNA Generated following Lariat Debranching Mediates Targeting of AID to DNA. *Cell* **161**, 762–773 (2015).

125. Chaudhuri, J. *et al.* Transcription-targeted DNA deamination by the AID antibody diversification enzyme. *Nature* **422**, 726–730 (2003).

126. Vohhodina, J. *et al.* BRCA1 binds TERRA RNA and suppresses R-Loop-based telomeric DNA damage. *Nat. Commun.* **12**, 1–16 (2021).

127. Erdel, F. *et al.* Telomere Recognition and Assembly Mechanism of Mammalian Shelterin. *Cell Rep.* **18**, 41–53 (2017).

128. Ducray, C., Pommier, J. P., Martins, L., Boussin, F. D. & Sabatier, L. Telomere dynamics, end-to-end fusions and telomerase activation during the human fibroblast immortalization process.

Oncogene **18**, 4211–4223 (1999).

129. Van Steensel, B., Smogorzewska, A. & De Lange, T. TRF2 protects human telomeres from end-to-end fusions. *Cell* **92**, 401–413 (1998).

130. Wynford-Thomas, D. & Kipling, D. The end-replication problem. *Nature* **389**, 551 (1997).

131. Abdallah, P. *et al.* A two-step model for senescence triggered by a single critically short telomere. *Nat. Cell Biol.* **11**, 988–993 (2009).

132. Karlseder, J., Smogorzewska, A. & De Lange, T. Senescence induced by altered telomere state, not telomere loss. *Science (80-.).* **295**, 2446–2449 (2002).

133. Hiyama, E. & Hiyama, K. Telomere and telomerase in stem cells. *Br. J. Cancer* **96**, 1020–1024 (2007).

134. Ozturk, S. Telomerase Activity and Telomere Length in Male Germ Cells. *Biol. Reprod.* **92**, 1-11,53 (2015).

135. Bekaert, S., Derradji, H. & Baatout, S. Telomere biology in mammalian germ cells and during development. *Dev. Biol.* **274**, 15–30 (2004).

136. Shay, J. W. & Bacchetti, S. of Telomerase Activity in Human Cancer. **33**, (1997).

137. Kim, N. W. *et al.* Specific association of human telomerase activity with immortal cells and cancer. *Science (80-.).* **266**, 2011–2015 (1994).

138. Dilley, R. L. & Greenberg, R. A. ALTernative Telomere Maintenance and Cancer. *Trends in cancer* **1**, 145–156 (2015).

139. Min, J., Wright, W. E. & Shay, J. W. Alternative lengthening of telomeres can be maintained by preferential elongation of lagging strands. *Nucleic Acids Res.* **45**, 2615–2628 (2017).

140. Rippe, K. & Luke, B. TERRA and the state of the telomere. *Nat. Struct. Mol. Biol.* **22**, 853–858 (2015).

141. Porro, A., Feuerhahn, S., Reichenbach, P. & Lingner, J. Molecular Dissection of Telomeric Repeat-Containing RNA Biogenesis Unveils the Presence of Distinct and Multiple Regulatory Pathways. *Mol. Cell. Biol.* **30**, 4808–4817 (2010).

142. Arora, R. *et al.* RNaseH1 regulates TERRA-telomeric DNA hybrids and telomere maintenance in ALT tumour cells. *Nat. Commun.* **5**, 1–11 (2014).

143. Silva, B. *et al.* FANCM limits ALT activity by restricting telomeric replication stress induced by deregulated BLM and R-loops. *Nat. Commun.* **10**, 1–16 (2019).

144. Skourti-Stathaki, K., Proudfoot, N. J. & Gromak, N. Human Senataxin Resolves RNA/DNA Hybrids Formed at Transcriptional Pause Sites to Promote Xrn2-Dependent Termination. *Mol. Cell* **42**, 794–805 (2011).

145. West, S. & Proudfoot, N. J. Transcriptional Termination Enhances Protein Expression in Human Cells. *Mol. Cell* **33**, 354–364 (2009).

146. Proudfoot, N. J. Transcriptional translation in mammals: Stopping de RNA polymerase II juggernaut. *Science (80-.).* **352**, 1–22 (2016).

147. West, S., Proudfoot, N. J. & Dye, M. J. Molecular Dissection of Mammalian RNA Polymerase II Transcriptional Termination. *Mol. Cell* **29**, 600–610 (2008).

148. West, S., Gromak, N. & Proudfoot, N. J. Human 5′ → 3′ exonuclease Xm2 promotes transcription termination at co-transcriptional cleavage sites. *Nature* **432**, 522–525 (2004).

149. Gromak, N., West, S. & Proudfoot, N. J. Pause Sites Promote Transcriptional Termination of Mammalian RNA Polymerase II. *Mol. Cell. Biol.* **26**, 3986–3996 (2006).

150. Smolka, J. A., Sanz, L. A., Hartono, S. R. & Chédin, F. Recognition of cellular RNAs by the S9.6 antibody creates pervasive artefacts when imaging RNA:DNA hybrids. *bioRxiv* 2020.01.11.902981 (2020) doi:10.1101/2020.01.11.902981.

151. Sanz, L. A. & Chédin, F. High-resolution, strand-specific R-loop mapping via S9.6-based DNA–RNA immunoprecipitation and high-throughput sequencing. *Nat. Protoc.* **14**, 1734–1755 (2019).

152. Cristini, A., Groh, M., Kristiansen, M. S. & Gromak, N. RNA/DNA Hybrid Interactome Identifies DXH9 as a Molecular Player in Transcriptional Termination and R-Loop-Associated DNA Damage. *Cell Rep.* **23**, 1891–1905 (2018).

153. Proudfoot, N. J. Ending the message: Poly(A) signals then and now. *Genes Dev.* **25**, 1770–1782 (2011).

154. Kireeva, M. L., Komissarova, N. & Kashlev, M. Overextended RNA:DNA hybrid as a negative regulator of RNA polymerase II processivity. *J. Mol. Biol.* **299**, 325–335 (2000).

155. Skourti-Stathaki, K., Kamieniarz-Gdula, K. & Proudfoot, N. J. R-loops induce repressive chromatin marks over mammalian gene terminators. *Nature* **516**, 436–439 (2014).

156. Yan, Q., Shields, E. J., Bonasio, R. & Sarma, K. Mapping Native R-Loops Genome-wide Using a Targeted Nuclease Approach. *Cell Rep.* **29**, 1369-1380.e5 (2019).

157. Chédin, F., Ginno, P. A., Lott, P. L., Christensen, H. C. & Korf, I. R-Loop Formation Is a Distinctive Characteristic of Unmethylated Human CpG Island Promoters. *Mol. Cell* **45**, 814–825 (2012).

158. Baylin, S. B. DNA methylation and gene silencing in cancer. *Nat. Clin. Pract. Oncol.* **2**, S4–S11 (2005).

159. Newell-Price, J., Clark, A. J. & King, P. DNA methylation and silencing of gene expression. *Trends Endocrinol. Metab.* **11**, 142–148 (2000).

160. Namba, S. *et al.* Differential regulation of CpG island methylation within divergent and unidirectional promoters in colorectal cancer. *Cancer Sci.* **110**, 1096–1104 (2019).

161. Rice, J. C., Ozcelik, H., Maxeiner, P., Andrulis, I. & Futscher, B. W. Methylation of the BRCA1 promoter is associated with decreased BRCA1 mRNA levels in clinical breast cancer specimens. *Carcinogenesis* **21**, 1761–1765 (2000).

162. Herman, J. G. *et al.* Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 6870–6875 (1998).

163. Grunseich, C. *et al.* Senataxin Mutation Reveals How R-Loops Promote Transcription by Blocking DNA Methylation at Gene Promoters. *Mol. Cell* **69**, 426-437.e7 (2018).

164. Arab, K. *et al.* GADD45A binds R-loops and recruits TET1 to CpG island promoters. *Nat. Genet.* **51**, 217–223 (2019).

165. Boque-Sastre, R. *et al.* Head-to-head antisense transcription and R-loop formation promotes transcriptional activation. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 5785–5790 (2015).

166. Ohle, C. *et al.* Transient RNA-DNA Hybrids Are Required for Efficient Double-Strand Break Repair. *Cell* **167**, 1001-1013.e7 (2016).

167. Pessina, F. *et al.* Functional transcription promoters at DNA double-strand breaks mediate RNA-driven phase separation of damage-response factors. *Nat. Cell Biol.* **21**, 1286–1299 (2019).

168. Marnef, A. & Legube, G. R-loops as Janus-faced modulators of DNA repair. *Nat. Cell Biol.* **23**, 305–313 (2021).

169. Britton, S. *et al.* DNA damage triggers SAF-A and RNA biogenesis factors exclusion from

chromatin coupled to R-loops removal. *Nucleic Acids Res.* **42**, 9047–9062 (2014).

170. Zhang, C. *et al.* METTL3 and N6-Methyladenosine Promote Homologous Recombination-Mediated Repair of DSBs by Modulating DNA-RNA Hybrid Accumulation. *Mol. Cell* **79**, 425-442.e7 (2020).

171. Teng, Y. *et al.* ROS-induced R loops trigger a transcription-coupled but BRCA1/2-independent homologous recombination pathway through CSB. *Nat. Commun.* **9**, 4115 (2018).

172. Hamperl, S., Bocek, M. J., Saldivar, J. C., Swigut, T. & Cimprich, K. A. Transcription-Replication Conflict Orientation Modulates R-Loop Levels and Activates Distinct DNA Damage Responses. *Cell* **170**, 774-786.e19 (2017).

173. Iannelli, F. *et al.* A damaged genome's transcriptional landscape through multilayered expression profiling around in situ-mapped DNA double-strand breaks. *Nat. Commun.* **8**, 1–7 (2017).

174. Collin, P., Jeronimo, C., Poitras, C. & Robert, F. RNA Polymerase II CTD Tyrosine 1 Is Required for Efficient Termination by the Nrd1-Nab3-Sen1 Pathway. *Mol. Cell* **73**, 655-669.e7 (2019).

175. Burger, K., Schlackow, M. & Gullerova, M. Tyrosine kinase c-Abl couples RNA polymerase II transcription to DNA double-strand breaks. *Nucleic Acids Res.* **47**, 3467–3484 (2019).

176. Shah, N. *et al.* Tyrosine-1 of RNA Polymerase II CTD Controls Global Termination of Gene Transcription in Mammals. *Mol. Cell* **69**, 48-61.e6 (2018).

177. Awwad, S. W., Abu-Zhayia, E. R., Guttmann-Raviv, N. & Ayoub, N. NELF -E is recruited to DNA double-strand break sites to promote transcriptional repression and repair . *EMBO Rep.* **18**, 745–764 (2017).

178. Gong, F., Clouaire, T., Aguirrebengoa, M., Legube, G. & Miller, K. M. Histone demethylase KDM5A regulates the ZMY ND8-NuRD chromatin remodeler to promote DNA repair. *J. Cell Biol.* **216**, 1959–1974 (2017).

179. Shanbhan, N. M., Rafalska-Metcalf, I. U., Balane- Bolivar, C., Janicki, S. M. & Greenberg, R. A. An ATM-Dependent Transcriptional Silencing Program is Transmitted Through Chromatin in Cis to DNA Double Strand Breaks. *Cell* **141**, 970–981 (2011).

180. Savitsky, P. *et al.* Multivalent Histone and DNA Engagement by a PHD/BRD/PWWP Triple Reader Cassette Recruits ZMYND8 to K14ac-Rich Chromatin. *Cell Rep.* **17**, 2724–2737 (2016).

181. Sharma, S. *et al.* MRE11-RAD50-NBS1 Complex Is Sufficient to Promote Transcription by RNA Polymerase II at Double-Strand Breaks by Melting DNA Ends. *Cell Rep.* **34**, 108565 (2021).

182. D'Alessandro, G. *et al.* BRCA2 controls DNA:RNA hybrid level at DSBs by mediating RNase H2 recruitment. *Nat. Commun.* **9**, 5376 (2018).

183. Sessa, G. *et al.* BRCA2 promotes DNA-RNA hybrid resolution by DDX5 helicase at DNA breaks to facilitate their repair‡. *EMBO J.* **40**, 1–25 (2021).

184. Dasgupta, S., Masukata, H. & Tomizawa, J. ichi. Multiple mechanisms for initiation of ColE1 DNA replication: DNA synthesis in the presence and absence of ribonuclease H. *Cell* **51**, 1113–1122 (1987).

185. Itoh, T. & Tomizawa, J. Formation of an RNA primer for initiation of replication of ColE1 DNA by ribonuclease H. *Proc. Natl. Acad. Sci. U. S. A.* **77**, 2450–2454 (1980).

186. Marians, K. J. Prokaryotic DNA replication. *Annu. Rev. Biochem.* **61**, 673–719 (1992).

187. Carles-Kinch, K. & Kreuzer, K. N. RNA-DNA hybrid formation at a bacteriophage T4 replication origin. *J. Mol. Biol.* **266**, 915–926 (1997).

188. Lombraña, R., Almeida, R., Álvarez, A. & Gómez, M. R-loops and initiation of DNA replication in human cells: A missing link? *Front. Genet.* **6**, 1–7 (2015).

189. Clayton, D. A. Replication of animal mitochondrial DNA. *Cell* **28**, 693–705 (1982).

190. Xu, B. & Clayton, D. A. A persistent RNA-DNA hybrid is formed during transcription at a phylogenetically conserved mitochondrial DNA sequence. *Mol. Cell. Biol.* **15**, 580–589 (1995).

191. Sequeira-Mendes, J. *et al.* Transcription initiation activity sets replication origin efficiency in mammalian cells. *PLoS Genet.* **5**, (2009).

192. Hoshina, S. *et al.* Human origin recognition complex binds preferentially to G-quadruplex-preferable RNA and single-stranded DNA. *J. Biol. Chem.* **288**, 30161–30171 (2013).

193. Eladl, A. *et al.* Investigation of the interaction of human origin recognition complex subunit 1 with g-quadruplex dnas of human c-myc promoter and telomere regions. *Int. J. Mol. Sci.* **22**, (2021).

194. Cayrou, C. *et al.* Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features. *Genome Res.* **21**, 1438–1449 (2011).

195. Cayrou, C. *et al.* New insights into replication origin characteristics in metazoans. *Cell Cycle* **11**, 658–667 (2012).

196. Prorok, P. *et al.* Involvement of G-quadruplex regions in mammalian replication origin activity. *Nat. Commun.* **10**, 1–16 (2019).

197. Valton, A. L. *et al.* G4 motifs affect origin positioning and efficiency in two vertebrate replicators. *EMBO J.* **33**, 732–746 (2014).

198. Chaudhuri, A. R. *et al.* Replication fork stability confers chemoresistance in BRCA-deficient cells. *Nature* **535**, 382–387 (2016).

199. Rondinelli, B. *et al.* EZH2 promotes degradation of stalled replication forks by recruiting MUS81 through histone H3 trimethylation. *Nat. Cell Biol.* **19**, 1371–1378 (2017).

200. Di Marco, S. *et al.* RECQ5 Helicase Cooperates with MUS81 Endonuclease in Processing Stalled Replication Forks at Common Fragile Sites during Mitosis. *Mol. Cell* **66**, 658-671.e8 (2017).

201. Matos, D. A. *et al.* ATR Protects the Genome against R Loops through a MUS81-Triggered Feedback Loop. *Mol. Cell* **77**, 514-527.e4 (2020).

202. Vilette, D., Ehrlich, S. D. & Michel, B. Transcription-induced deletions in plasmid vectors: M13 DNA replication as a source of instability. *Mol. Gen. Genet.* **252**, 398–403 (1996).

203. Srivatsan, A., Tehranchi, A., MacAlpine, D. M. & Wang, J. D. Co-orientation of replication and transcription preserves genome integrity. *PLoS Genet.* **6**, (2010).

204. Merrikh, H., Zhang, Y., Grossman, A. D. & Wang, J. D. Replication-transcription conflicts in bacteria. *Nat. Rev. Microbiol.* **10**, 449–458 (2012).

205. Rocha, E. P. Evolutionary patterns in prokaryotic genomes. *Curr. Opin. Microbiol.* **11**, 454–460 (2008).

206. Aguilera, A. The connection between transcription and genomic instability. *EMBO J.* **21**, 195–201 (2002).

207. Sollier, J. *et al.* Transcription-Coupled Nucleotide Excision Repair Factors Promote R-Loop-Induced Genome Instability. *Mol. Cell* **56**, 777–785 (2014).

208. Cristini, A. *et al.* Dual Processing of R-Loops and Topoisomerase I Induces Transcription-Dependent DNA Double-Strand Breaks. *Cell Rep.* **28**, 3167-3181.e6 (2019).

209. Chakraborty, P., Huang, J. T. J. & Hiom, K. DHX9 helicase promotes R-loop formation in cells with impaired RNA splicing. *Nat. Commun.* **9**, (2018).

210. Chen, L. *et al.* The Augmented R-Loop Is a Unifying Mechanism for Myelodysplastic Syndromes Induced by High-Risk Splicing Factor Mutations. *Mol. Cell* **69**, 412-425.e6 (2018).

211. Ogawa, S. Splicing factor mutations in myelodysplasia. *Int. J. Hematol.* **96**, 438–442 (2012).

212. Pellagatti, A. & Boultwood, J. Splicing factor gene mutations in the myelodysplastic syndromes: impact on disease phenotype and therapeutic applications. *Adv. Biol. Regul.* **63**, 59–70 (2017).

213. Boultwood, J., Dolatshad, H., Varanasi, S. S., Yip, B. H. & Pellagatti, A. The role of splicing factor mutations in the pathogenesis of the myelodysplastic syndromes. *Adv. Biol. Regul.* **54**, 153–161 (2014).

214. Wang, I. X. *et al.* Human proteins that interact with RNA/DNA hybrids. *Genome Res.* **28**, 1405–1414 (2018).

215. Song, C., Hotz-wagenblatt, A., Voit, R. & Grummt, I. SIRT7 and the DEAD-box helicase DDX21 cooperate to resolve genomic R loops and safeguard genome stability. 1370–1381 (2017) doi:10.1101/gad.300624.117.1370.

216. Mersaoui, S. Y. *et al.* Arginine methylation of the DDX 5 helicase RGG / RG motif by PRMT 5 regulates resolution of RNA:DNA hybrids . *EMBO J.* **38**, 1–20 (2019).

217. Hodroj, D. *et al.* An ATR -dependent function for the Ddx19 RNA helicase in nuclear R-loop metabolism . *EMBO J.* **36**, 1182–1198 (2017).

218. Talwar, T. *et al.* The DEAD-box protein DDX43 (HAGE) is a dual RNA-DNA helicase and has a K-homology domain required for full. **43**, 10429–10443 (2017).

219. Pérez-Calero, C. *et al.* UAP56/DDX39B is a major cotranscriptional RNA–DNA helicase that unwinds harmful R loops genome-wide. *Genes Dev.* **34**, 1–15 (2020).

220. Yuan, W. *et al.* TDRD3 promotes DHX9 chromatin recruitment and R-loop resolution. 1–19 (2021).

221. Chakraborty, P. & Hiom, K. DHX9-dependent recruitment of BRCA1 to RNA is required to promote DNA end resection in homologous recombination. *bioRxiv* **1**, 2019.12.20.884593 (2019).

222. Wahba, L., Amon, J. D., Koshland, D. & Vuica-Ross, M. RNase H and Multiple RNA Biogenesis Factors Cooperate to Prevent RNA:DNA Hybrids from Generating Genome Instability. *Mol. Cell* **44**, 978–988 (2011).

223. Cerritelli, S. M. & Crouch, R. J. Ribonuclease H: The enzymes in eukaryotes. *FEBS J.* **276**, 1494–1505 (2009).

224. Stein, H. & Hausen, P. Enzyme from calf thymus degrading the RNA moiety of DNA-RNA Hybrids: effect on DNA-dependent RNA polymerase. *Science* **166**, 393–395 (1969).

225. Nowotny, M. *et al.* Structure of Human RNase H1 Complexed with an RNA/DNA Hybrid: Insight into HIV Reverse Transcription. *Mol. Cell* **28**, 264–276 (2007).

226. Nowotny, M. *et al.* Specific recognition of RNA/DNA hybrid and enhancement of human RNase H1 activity by HBD. *EMBO J.* **27**, 1172–1181 (2008).

227. Nguyen, H. D. *et al.* Functions of Replication Protein A as a Sensor of R Loops and a Regulator of RNaseH1. *Mol. Cell* **65**, 832-847.e4 (2017).

228. Cerritelli, S. M. *et al.* Failure to Produce Mitochondrial DNA Results in Embryonic Lethality in Rnaseh1 Null Mice strand mode of DNA synthesis and that the RNA-DNA hybrids are RNA primers of DNA replication. The pres-ence of abundant RNA-DNA hybrids in replicating Laboratory of M. *Mol. Cell* **11**, 807–815 (2003).

229. Akman, G. *et al.* Pathological ribonuclease H1 causes R-loop depletion and aberrant DNA segregation in mitochondria. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E4276–E4285 (2016).

230. Parajuli, S. *et al.* Human ribonuclease H1 resolves R-loops and thereby enables progression of the DNA replication fork. *J. Biol. Chem.* **292**, 15216–15224 (2017).

231. Zhao, H., Zhu, M., Limbo, O. & Russell, P. RNase H eliminates R-loops that disrupt DNA replication but is nonessential for efficient DSB repair. *EMBO Rep.* **19**, 1–10 (2018).

232. Crow, Y. J. *et al.* Mutations in genes encoding ribonuclease H2 subunits cause Aicardi-Goutières syndrome and mimic congenital viral brain infection. *Nat. Genet.* **38**, 910–916 (2006).

233. Eder, P. S., Walder, R. Y. & Walder, J. A. Substrate specificity of human RNase H1 and its role in excision repair of ribose residues misincorporated in DNA. *Biochimie* **75**, 123–126 (1993).

234. Serrano Cardona, L. & Muñoz Mata, E. RNase H2-Initiated Ribonucleotide Excision Repair. *Early Hum. Dev.* **83**, 1–11 (2013).

235. Hiller, B. *et al.* Mammalian RNase H2 removes ribonucleotides from DNA to maintain genome integrity. *J. Exp. Med.* **209**, 1419–1426 (2012).

236. Liu, B., Hu, J., Wang, J. & Kong, D. Direct visualization of RNA-DNA primer removal from okazaki fragments provides support for flap cleavage and exonucleolytic pathways in eukaryotic cells. *J. Biol. Chem.* **292**, 4777–4788 (2017).

237. Lockhart, A. *et al.* RNase H1 and H2 Are Differentially Regulated to Process RNA-DNA Hybrids. *Cell Rep.* **29**, 2890-2900.e5 (2019).

238. Marnef, A. & Legube, G. m6A RNA modification as a new player in R-loop regulation. *Nat. Genet.* **52**, 27–28 (2020).

239. Yang, X. *et al.* m6A promotes R-loop formation to facilitate transcription termination. *Cell Res.* **29**, 1035–1038 (2019).

240. Abakir, A. *et al.* N 6-methyladenosine regulates the stability of RNA:DNA hybrids in human cells. *Nat. Genet.* **52**, 48–55 (2020).

241. Kang, H. J. *et al.* TonEBP recognizes R-loops and initiates m6A RNA methylation for R-loop resolution. *Nucleic Acids Res.* **49**, 269–284 (2021).

242. Yan, Q. & Sarma, K. MapR: A Method for Identifying Native R-Loops Genome Wide. *Curr. Protoc. Mol. Biol.* **130**, 1–12 (2020).

243. Negrini, S., Gorgoulis, V. G. & Halazonetis, T. D. Genomic instability--an evolving hallmark of cancer. *Nat. Rev. Mol. Cell Biol.* **11**, 220–228 (2010).

244. Alsop, K. *et al.* BRCA mutation frequency and patterns of treatment response in BRCA mutation-positive women with ovarian cancer: A report from the Australian ovarian cancer study group. *J. Clin. Oncol.* **30**, 2654–2663 (2012).

245. Manchana, T., Phoolcharoen, N. & Tantbirojn, P. BRCA mutation in high grade epithelial ovarian cancers. *Gynecol. Oncol. Reports* **29**, 102–105 (2019).

246. Kuchenbaecker, K. B. *et al.* Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. *JAMA - J. Am. Med. Assoc.* **317**, 2402–2416 (2017).

247. Hatchi, E. *et al.* BRCA1 Recruitment to Transcriptional Pause Sites Is Required for R-Loop-Driven DNA Damage Repair. *Mol. Cell* **57**, 636–647 (2015).

248. Richard, P. & Manley, J. L. Orexin activation counteracts decreases in nonexercise activity thermogenesis (NEAT) caused by high-fat diet. *J Mol Biol.* **429**, 3168–3180 (2017).

249. Moreira, M. C. *et al.* Senataxin, the ortholog of a yeast RNA helicase, is mutant in ataxia-ocular apraxia 2. *Nat. Genet.* **36**, 225–227 (2004).

250. Chen, Y. Z. *et al.* DNA/RNA helicase gene mutations in a form of juvenile amyotrophic lateral sclerosis (ALS4). *Am. J. Hum. Genet.* **74**, 1128–1135 (2004).

251. Tefferi, A. & Vardiman, J. W. Myelodysplastic Syndromes. *N. Engl. J. Med.* **361**, 1872–1885 (2009).

252. Kim, E. *et al.* SRSF2 Mutations Contribute to Myelodysplasia by Mutant-Specific Effects on Exon Recognition. *Cancer Cell* **27**, 617–630 (2015).

253. Ilagan, J. O. *et al.* U2AF1 mutations alter splice site recognition in hematological malignancies. *Genome Res.* **25**, 14–26 (2015).

254. Zhou, T. *et al.* Myelodysplastic syndrome: An inability to appropriately respond to damaged DNA? *Exp. Hematol.* **41**, 665–674 (2013).

255. Chen, L. *et al.* R-ChIP Using Inactive RNase H Reveals Dynamic Coupling of R-loops with Transcriptional Pausing at Gene Promoters. *Mol. Cell* **68**, 745-757.e5 (2017).

256. Wang, K. *et al.* Genomic profiling of native R loops with a DNA-RNA hybrid recognition sensor. 1–17 (2021).

257. Malig, M., Hartono, S. R., Giafaglione, J. M., Sanz, L. A. & Chedin, F. Ultra-deep Coverage Single-molecule R-loop Footprinting Reveals Principles of R-loop Formation. *J. Mol. Biol.* **432**, 2271–2288 (2020).

258. Wissink, E. M., Vihervaara, A., Tippens, N. D. & Lis, J. T. Nascent RNA analyses: tracking transcription and its regulation. *Nat. Rev. Genet.* **20**, 705–723 (2019).

259. Henriques, T. *et al.* Stable pausing by rna polymerase II provides an opportunity to target and integrate regulatory signals. *Mol. Cell* **52**, 517–528 (2013).

260. Weinreb, J. T. *et al.* Excessive R-loops trigger an inflammatory cascade leading to increased HSPC production Article Excessive R-loops trigger an inflammatory cascade leading to increased HSPC production. *Dev. Cell* 1–14 (2021) doi:10.1016/j.devcel.2021.02.006.

261. Liu, Y. & Imai, R. Function of plant DExD/H-Box RNA helicases associated with ribosomal RNA biogenesis. *Front. Plant Sci.* **9**, 1–7 (2018).

262. Tsukamoto, T. *et al.* Insights into the Involvement of Spliceosomal Mutations in Myelodysplastic Disorders from Analysis of SACY-1/DDX41 in Caenorhabditis elegans . *Genetics* **214**, 869–893 (2020).

263. Jiang, Y., Zhu, Y., Liu, Z. J. & Ouyang, S. The emerging roles of the DDX41 protein in immunity and diseases. *Protein Cell* **8**, 83–89 (2017).

264. Maciejewski, J. P., Padgett, R. A., Brown, A. L. & Müller-Tidow, C. DDX41-related myeloid neoplasia. *Semin. Hematol.* **54**, 94–97 (2017).

265. Kadono, M. *et al.* Biological implications of somatic DDX41 p . R525H mutation in acute myeloid leukemia. *Exp. Hematol.* **44**, 745-754.e4 (2016).

266. Omura, H. *et al.* Structural and functional analysis of DDX41: A bispecific immune receptor for DNA and cyclic dinucleotide. *Sci. Rep.* **6**, 1–4 (2016).

267. Jiang, Y. *et al.* Structural and functional analyses of human DDX41 DEAD domain. *Protein Cell* **8**, 72–76 (2017).

268. Putnam, A. A. & Jankowsky, E. DEAD-box helicases as integrators of RNA, nucleotide and protein binding. *Biochim. Biophys. Acta* **1829**, 884–893 (2013).

269. Cheah, J. J. C., Hahn, C. N., Hiwase, D. K., Scott, H. S. & Brown, A. L. Myeloid neoplasms with germline DDX41 mutation. *Int. J. Hematol.* **106**, 163–174 (2017).

270. Polprasert, C. *et al.* Inherited and Somatic Defects in DDX41 in Myeloid Neoplasms. *Cancer Cell* **27**, 658–670 (2015).

271. Álvarez-Quilón, A. *et al.* ATM specifically mediates repair of double-strand breaks with blocked DNA ends. *Nat. Commun.* **5**, 3347 (2014).

272. Rondeau, S. *et al.* ATM has a major role in the double-strand break repair pathway dysregulation in sporadic breast carcinomas and is an independent prognostic marker at both mRNA and protein levels. *Br. J. Cancer* **112**, 1059–1066 (2015).

273. Zannini, L., Delia, D. & Buscemi, G. CHK2 kinase in the DNA damage response and beyond. *J. Mol. Cell Biol.* **6**, 442–457 (2014).

274. Buscemi, G. *et al.* Activation of ATM and Chk2 kinases in relation to the amount of DNA strand breaks. *Oncogene* **23**, 7691–7700 (2004).

275. Marabitti, V. *et al.* ATM pathway activation limits R-loop-associated genomic instability in Werner syndrome cells. *Nucleic Acids Res.* **47**, 3485–3502 (2019).

276. Tresini, M. *et al.* The core spliceosome as target and effector of non-canonical ATM signalling. *Nature* **523**, 53–58 (2015).

277. Gupta, D., Lin, B., Cowan, A. & Heinen, C. D. ATR-Chk1 activation mitigates replication stress caused by mismatch repair-dependent processing of DNA damage. *Proc. Natl. Acad. Sci. U. S. A.* **115**, 1523–1528 (2018).

278. Flynn, R. L. & Zou, L. ATR: a master conductor of cellular responses to DNA replication stress.

Trends Biochem. Sci. **36**, 133–140 (2011).

279. Lee, Y.-C., Zhou, Q., Chen, J. & Yuan, J. RPA-Binding Protein ETAA1 Is an ATR Activator Involved in DNA Replication Stress Response. *Curr. Biol.* **26**, 3257–3268 (2016).

280. Barroso, S. *et al.* The DNA damage response acts as a safeguard against harmful DNA–RNA hybrids of different origins. *EMBO Rep.* **20**, (2019).

281. Ragu, S., Matos-Rodrigues, G. & Lopez, B. S. Replication stress, DNA damage, inflammatory cytokines and innate immune response. *Genes (Basel).* **11**, (2020).

282. Newton, K. & Dixit, V. M. Signaling in innate immunity and inflammation. *Cold Spring Harb. Perspect. Biol.* **4**, (2012).

283. Paludan, S. R. & Bowie, A. G. Immune Sensing of DNA. *Immunity* **38**, 870–880 (2013).

284. Roers, A., Hiller, B. & Hornung, V. Recognition of Endogenous Nucleic Acids by the Innate Immune System. *Immunity* **44**, 739–754 (2016).

285. Gasser, S. & Raulet, D. H. The DNA damage response arouses the immune system. *Cancer Res.* **66**, 3959–3962 (2006).

286. Li, T. & Chen, Z. J. The cGAS-cGAMP-STI NG pathway connects DNA damage to inflammation, senescence, and cancer. *J. Exp. Med.* **215**, 1287–1299 (2018).

287. Beli, P. *et al.* Proteomic Investigations Reveal a Role for RNA Processing Factor THRAP3 in the DNA Damage Response. *Mol. Cell* **46**, 212–225 (2012).

288. Gasser, S. *et al.* Sensing of dangerous DNA. *Mech. Ageing Dev.* **165**, 33–46 (2017).

289. Civril, F. *et al.* Structural mechanism of cytosolic DNA sensing by cGAS. *Nature* **498**, 332–337 (2013).

290. Luecke, S. *et al.* cGAS is activated by DNA in a length-dependent manner . *EMBO Rep.* **18**, 1707–1715 (2017).

291. Motwani, M., Pesiridis, S. & Fitzgerald, K. A. DNA sensing by the cGAS–STING pathway in health and disease. *Nat. Rev. Genet.* **20**, 657–674 (2019).

292. Shang, G., Zhang, C., Chen, Z. J., Bai, X. chen & Zhang, X. Cryo-EM structures of STING reveal its mechanism of activation by cyclic GMP–AMP. *Nature* **567**, 389–393 (2019).

293. Zhang, C. *et al.* Structural basis of STING binding with and phosphorylation by TBK1. *Nature*

567, 394–398 (2019).

294. Ishikawa, H., Ma, Z. & Barber, G. N. STING regulates intracellular DNA-mediated, type i interferon-dependent innate immunity. *Nature* **461**, 788–792 (2009).

295. Yasuo, T. & J., C. Z. STING Specifies IRF3 Phosphorylation by TBK1 in the Cytosolic DNA Signaling Pathway. *Sci. Signal.* **5**, ra20–ra20 (2012).

296. Yum, S., Li, M., Fang, Y. & Chen, Z. J. TBK1 recruitment to STING activates both IRF3 and NF-κB that mediate immune defense against tumors and viral infections. *Proc. Natl. Acad. Sci. U. S. A.* **118**, 1–9 (2021).

297. Rongtuan, L., Christophe, H., M., P. P. & John, H. Virus-Dependent Phosphorylation of the IRF-3 Transcription Factor Regulates Nuclear Translocation, Transactivation Potential, and Proteasome-Mediated Degradation. *Mol. Cell. Biol.* **18**, 2986–2996 (1998).

298. Zhu, M., Fang, T., Li, S., Meng, K. & Guo, D. Bipartite Nuclear Localization Signal Controls Nuclear Import and DNA-Binding Activity of IFN Regulatory Factor 3. *J. Immunol.* **195**, 289–297 (2015).

299. Möser, C. V. *et al.* TANK-binding kinase 1 (TBK1) modulates inflammatory hyperalgesia by regulating MAP kinases and NF-κB dependent genes. *J. Neuroinflammation* **12**, 1–13 (2015).

300. Buss, H. *et al.* Constitutive and interleukin-1-inducible phosphorylation of p65 NF-κB at serine 536 is mediated by multiple protein kinases including IκB kinase (IKK)-α, IKKβ, IKKε, TRAF family member-associated (TANK)-binding kinase 1 (TBK1), and an unknown kinase and c. *J. Biol. Chem.* **279**, 55633–55643 (2004).

301. MacKenzie, K. J. *et al.* CGAS surveillance of micronuclei links genome instability to innate immunity. *Nature* **548**, 461–465 (2017).

302. Kwon, M., Leibowitz, M. L. & Lee, J. H. Small but mighty: the causes and consequences of micronucleus rupture. *Exp. Mol. Med.* **52**, 1777–1786 (2020).

303. Cinat, D., Coppes, R. P. & Barazzuol, L. DNA Damage-Induced Inflammatory Microenvironment and Adult Stem Cell Response. *Front. Cell Dev. Biol.* **9**, 1–18 (2021).

304. Jahandideh, B. *et al.* The pro-Inflammatory cytokines effects on mobilization, self-renewal and differentiation of hematopoietic stem cells. *Hum. Immunol.* **81**, 206–217 (2020).

305. Hormaechea-Agulla, D., Le, D. T. & King, K. Y. Common Sources of Inflammation and Their Impact on Hematopoietic Stem Cell Biology. *Curr. Stem Cell Reports* **6**, 96–107 (2020).

306. Crop, M. J. *et al.* Inflammatory conditions affect gene expression and function of human adipose tissue-derived mesenchymal stem cells. *Clin. Exp. Immunol.* **162**, 474–486 (2010).

307. Lepperdinger, G. Inflammation and mesenchymal stem cell aging. *Curr. Opin. Immunol.* **23**, 518–524 (2011).

308. Zhang, Y., Ravikumar, M., Ling, L., Nurcombe, V. & Cool, S. M. Age-Related Changes in the Inflammatory Status of Human Mesenchymal Stem Cells: Implications for Cell Therapy. *Stem cell reports* **16**, 694–707 (2021).

309. Ambrosi, T. H. *et al.* Aged skeletal stem cells generate an inflammatory degenerative niche. *Nature* **597**, 256–262 (2021).

310. Caiado, F., Pietras, E. M. & Manz, M. G. Inflammation as a regulator of hematopoietic stem cell function in disease, aging, and clonal selection. *J. Exp. Med.* **218**, 1–19 (2021).

311. Jeong, Y. J., Oh, H. K., Park, S. H. & Bong, J. G. Association between inflammation and cancer stem cell phenotype in breast cancer. *Oncol. Lett.* **15**, 2380–2386 (2018).

312. Ng, K. W., Marshall, E. A., Bell, J. C. & Lam, W. L. cGAS-STING and Cancer: Dichotomous Roles in Tumor Immunity and Development. *Trends Immunol.* **39**, 44–54 (2018).

313. Chatzidoukaki, O. *et al.* R-loops trigger the release of cytoplasmic ssDNAs leading to chronic inflammation upon DNA damage. *Sci. Adv.* **7**, (2021).

314. Rees, J. S., Li, X.-W., Perrett, S., Lilley, K. S. & Jackson, A. P. Protein Neighbors and Proximity Proteomics. *Mol. Cell. Proteomics* **14**, 2848–2856 (2015).

315. Roux, K. J., Kim, D. I., Raida, M. & Burke, B. A promiscuous biotin ligase fusion protein identifies proximal and interacting proteins in mammalian cells. *J. Cell Biol.* **196**, 801–810 (2012).

316. Kim, D. I. *et al.* An improved smaller biotin ligase for BioID proximity labeling. *Mol. Biol. Cell* **27**, 1188–1196 (2016).

317. Branon, T. C. *et al.* Efficient proximity labeling in living cells and organisms with TurboID. *Nat. Biotechnol.* **36**, 880–898 (2018).

318. Martell, J. D. *et al.* Engineered ascorbate peroxidase as a genetically encoded reporter for electron microscopy. *Nat. Biotechnol.* **30**, 1143–1148 (2012).

319. Trinkle-Mulcahy, L. Recent advances in proximity-based labeling methods for interactome mapping [version 1; referees: 2 approved]. *F1000Research* **8**, (2019).

320. Lam, S. S. *et al.* Directed evolution of APEX2 for electron microscopy and proteomics. *Nat Methods* **12**, 51–54 (2015).

321. Lin, Q. *et al.* Screening of Proximal and Interacting Proteins in Rice Protoplasts by Proximity-Dependent Biotinylation. *Front. Plant Sci.* **8**, 749 (2017).

322. Dingar, D. *et al.* BioID identifies novel c-MYC interacting partners in cultured cells and xenograft tumors. *J. Proteomics* **118**, 95–111 (2015).

323. Gu, B., Lambert, J.-P., Cockburn, K., Gingras, A.-C. & Rossant, J. AIRE is a critical spindle-associated protein in embryonic stem cells. *Elife* **6**, (2017).

324. Antonicka, H. *et al.* A High-Density Human Mitochondrial Proximity Interaction Network. *Cell Metab.* **32**, 479-497.e9 (2020).

325. Rhee, H.-W. *et al.* Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. *Science* **339**, 1328–1331 (2013).

326. Hung, V. *et al.* Proteomic mapping of cytosol-facing outer mitochondrial and ER membranes in living human cells by proximity biotinylation. *Elife* **6**, 1–39 (2017).

327. Wolters, D. A., Washburn, M. P. & Yates, J. R. An Automated Multidimensional Protein Identification Technology for Shotgun Proteomics. *Anal. Chem.* **73**, 5683–5690 (2001).

328. Zhang, Y., Fonslow, B. R., Shan, B., Baek, M.-C. & Yates, J. R. Protein Analysis by Shotgun/Bottom-up Proteomics. *Chem. Rev.* **113**, 2343–2394 (2013).

329. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. **422**, (2003).

330. Olsen, J. V, Ong, S.-E. & Mann, M. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol. Cell. Proteomics* **3**, 608–614 (2004).

331. Tran, B. Q. *et al.* Addressing Trypsin Bias in Large Scale (Phospho)proteome Analysis by Size Exclusion Chromatography and Secondary Digestion of Large Post-Trypsin Peptides. *J. Proteome Res.* **10**, 800–811 (2011).

332. Manadas, B., Mendes, V. M., English, J. & Dunn, M. J. Peptide fractionation in proteomics approaches. *Expert Rev. Proteomics* **7**, 655–663 (2010).

333. Kebarle, P. & Verkerk, U. H. Electrospray: from ions in solution to ions in the gas phase, what we know now. *Mass Spectrom. Rev.* **28**, 898–917 (2009).

334. Loo, J. A., Udseth, H. R. & Smith, R. D. Peptide and protein analysis by electrospray ionization-

mass spectrometry and capillary electrophoresis-mass spectrometry. *Anal. Biochem.* **179**, 404–412 (1989).

335. Ho, C. S. *et al.* Electrospray ionisation mass spectrometry: principles and clinical applications. *Clin. Biochem. Rev.* **24**, 3–12 (2003).

336. Schröder, D. Coulomb explosions and stability of multiply charged ions in the gas phase. *Angew. Chemie - Int. Ed.* **43**, 1329–1331 (2004).

337. Steve, N. & B., F. J. Gas-phase ions of solute species from charged droplets of solutions. *Proc. Natl. Acad. Sci.* **104**, 1111–1117 (2007).

338. Iribarne, J. V. & Thomson, B. A. On the evaporation of small ions from charged droplets. *J. Chem. Phys.* **64**, 2287–2294 (1976).

339. Mellon, F. A. MASS SPECTROMETRY | Principles and Instrumentation. in (ed. Caballero, B. B. T.-E. of F. S. and N. (Second E.) 3739–3749 (Academic Press, 2003). doi:https://doi.org/10.1016/B0-12-227055-X/00746-X.

340. Miller, P. E. & Denton, M. B. The quadrupole mass filter: Basic operating concepts. *J. Chem. Educ.* **63**, 617–622 (1986).

341. Makarov, A. Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis. *Anal. Chem.* **72**, 1156–1162 (2000).

342. Hu, Q. *et al.* The Orbitrap: a new mass spectrometer. *J. Mass Spectrom.* **40**, 430–443 (2005).

343. Steen, H. & Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell Biol.* **5**, 699–711 (2004).

344. Hebert, A. S. *et al.* Improved Precursor Characterization for Data-Dependent Mass Spectrometry. *Anal. Chem.* **90**, 2333–2340 (2018).

345. Kalli, A., Smith, G. T., Sweredoski, M. J. & Hess, S. Evaluation and optimization of mass spectrometric settings during data-dependent acquisition mode: focus on LTQ-Orbitrap mass analyzers. *J. Proteome Res.* **12**, 3071–3086 (2013).

346. Guo, J. & Huan, T. Comparison of Full-Scan, Data-Dependent, and Data-Independent Acquisition Modes in Liquid Chromatography–Mass Spectrometry Based Untargeted Metabolomics. *Anal. Chem.* **92**, 8072–8080 (2020).

347. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-

range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).

348. Cox, J. *et al.* Accurate proteome-wide label-free quantification by delayed normalization and maximal peptide ratio extraction, termed MaxLFQ. *Mol. Cell. Proteomics* **13**, 2513–2526 (2014).

349. Zecha, J. *et al.* TMT Labeling for the Masses: A Robust and Cost-efficient, In-solution Labeling Approach. *Mol. Cell. Proteomics* **18**, 1468–1478 (2019).

350. Ong, S. E. & Mann, M. A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nat. Protoc.* **1**, 2650–2660 (2007).

351. Li, J. *et al.* TMTpro reagents: a set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples. *Nat. Methods* **17**, (2020).

352. Sandberg, A., Branca, R. M. M., Lehtiö, J. & Forshed, J. Quantitative accuracy in mass spectrometry based proteomics of complex samples: The impact of labeling and precursor interference. *J. Proteomics* **96**, 133–144 (2014).

353. Yu, Z. *et al.* DDX5 resolves R-loops at DNA double-strand breaks to promote DNA repair and avoid chromosomal deletions. *NAR Cancer* **2**, 1–19 (2020).

354. Gothe, H. J. *et al.* Spatial Chromosome Folding and Active Transcription Drive DNA Fragility and Formation of Oncogenic MLL Translocations. *Mol. Cell* **75**, 267-283.e12 (2019).

355. Bouwman, B. A. M. *et al.* Genome-wide detection of DNA double-strand breaks by in-suspension BLISS. *Nat. Protoc.* **15**, 3894–3941 (2020).

356. Chiarle, R. *et al.* Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. *Cell* **147**, 107–119 (2011).

357. Lensing, S. V, Marsico, G., Hänsel-hertsch, R. & Lam, E. Y. DSBCapture : in situ capture and direct sequencing of dsDNA breaks. *Nat Methods* **13**, 855–857 (2016).

358. Yan, W. X. *et al.* Breaks labeling in situ and sequencing (BLISS). *Protoc. Exch.* 1–15 (2017) doi:http://dx.doi.org/10.1038/protex.2017.018.

359. Elkon, R. *et al.* Myc coordinates transcription and translation to enhance transformation and suppress invasiveness. *EMBO Rep.* **16**, 1723–1736 (2015).

360. Klco, J. M. & Mullighan, C. G. Advances in germline predisposition to acute leukaemias and myeloid neoplasms. *Nat. Rev. Cancer* (2020) doi:10.1038/s41568-020-00315-z.

361. Chiang, H.-C. *et al.* BRCA1-associated R-loop affects transcription and differentiation in breast luminal epithelial cells. *Nucleic Acids Res.* **47**, 5086–5099 (2019).

362. Smolka, J. A., Sanz, L. A., Hartono, S. R. & Chédin, F. Recognition of RNA by the S9.6 antibody creates pervasive artifacts when imaging RNA:DNA hybrids. *J. Cell Biol.* **220**, (2021).

363. Richard, P. & Manley, J. L. R Loops and Links to Human Disease. *J. Mol. Biol.* **429**, 3168–3180 (2017).

364. Zhang, Y. *et al.* M6A modification in RNA: Biogenesis, functions and roles in gliomas. *J. Exp. Clin. Cancer Res.* **39**, 1–16 (2020).

365. Shin, J.-H. & Kelman, Z. The replicative helicases of bacteria, archaea, and eukarya can unwind RNA-DNA hybrid substrates. *J. Biol. Chem.* **281**, 26914–26921 (2006).

366. Kim, S. *et al.* ATAD5 restricts R-loop formation through PCNA unloading and RNA helicase maintenance at the replication fork. *Nucleic Acids Res.* **48**, 7218–7238 (2020).

367. Delgado, S., Bird, A. & Antequera, F. Initiation of DNA replication at CpG islands in mammalian chromosomes ´ a Go. **17**, 2426–2435 (1998).

368. Wang, X. *et al.* SMARCB1-mediated SWI/SNF complex function is essential for enhancer regulation. *Nat. Genet.* **49**, 289–295 (2017).

369. Nakayama, R. T. *et al.* SMARCB1 is required for widespread BAF complex-mediated activation of enhancers and bivalent promoters. *Nat. Genet.* **49**, 1613–1623 (2017).

370. Bayona-Feliu, A., Barroso, S., Muñoz, S. & Aguilera, A. The SWI/SNF chromatin remodeling complex helps resolve R-loop-mediated transcription-replication conflicts. *Nat. Genet.* **53**, 1050–1063 (2021).

371. Tsai, S. *et al.* ARID1A regulates R-loop associated DNA replication stress. *PLoS Genet.* **17**, e1009238 (2021).

372. Strzyz, P. R loops regulate chromatin remodelling. *Nat. Rev. Mol. Cell Biol.* **16**, 703 (2015).

373. Oppikofer, M. *et al.* Expansion of the ISWI chromatin remodeler family with new active complexes. *EMBO Rep.* **18**, 1697–1706 (2017).

374. Kang, H. J. *et al.* Thrap3 promotes R-loop resolution via interaction with methylated DDX5. *Exp. Mol. Med.* **53**, 1602–1611 (2021).

375. Villarreal, O. D., Mersaoui, S. Y., Yu, Z., Masson, J. Y. & Richard, S. Genome-wide R-loop

analysis defines unique roles for DDX5, XRN2, and PRMT5 in DNA/RNA hybrid resolution. *Life Sci. Alliance* **3**, 1–14 (2020).

376. Bader, A. S. *et al.* DDX17 is required for efficient DSB repair at DNA : RNA hybrid deficient loci. (2021).

377. Domínguez-Sánchez, M. S., Barroso, S., Gómez-González, B., Luna, R. & Aguilera, A. Genome instability and transcription elongation impairment in human cells depleted of THO/TREX. *PLoS Genet.* **7**, 19–22 (2011).

378. Stäßer, K. *et al.* TREX is a conserved complex coupling transcription with messenger RNA export. *Nature* **417**, 304–308 (2002).

379. Kellner, M. *et al.* DEAD-box helicase DDX27 regulates 3' end formation of ribosomal 47S RNA and stably associates with the PeBoW-complex. *Exp. Cell Res.* **334**, 146–159 (2015).

380. Choudhury, P., Hackert, P., Memet, I., Sloan, K. E. & Bohnsack, M. T. The human RNA helicase DHX37 is required for release of the U3 snoRNP from pre-ribosomal particles. *RNA Biol.* **16**, 54–68 (2019).

381. Pryor, A. *et al.* Growth-regulated expression and G0-specific turnover of the mRNA that encodes URH49, a mammalian DExH/D box protein that is highly related to the mRNA export protein UAP56. *Nucleic Acids Res.* **32**, 1857–1865 (2004).

382. Sugiura, T., Nagano, Y. & Noguchi, Y. DDX39, upregulated in lung squamous cell cancer, displays RNA helicase activities and promotes cancer cell growth. *Cancer Biol. Ther.* **6**, 957–964 (2007).

383. Uhlmann-Schiffler, H., Jalal, C. & Stahl, H. Ddx42p--a human DEAD box protein with RNA chaperone activities. *Nucleic Acids Res.* **34**, 10–22 (2006).

384. Paulsen, R. D. *et al.* A Genome-wide siRNA Screen Reveals Diverse Cellular Processes and Pathways that Mediate Genome Stability. *Mol. Cell* **35**, 228–239 (2009).

385. Shen, W. *et al.* Dynamic nucleoplasmic and nucleolar localization of mammalian RNase H1 in response to RNAP I transcriptional R-loops. *Nucleic Acids Res.* **45**, 10672–10692 (2017).

386. González de Cózar, J. M. *et al.* RNase H1 promotes replication fork progression through oppositely transcribed regions of Drosophila mitochondrial DNA. *J. Biol. Chem.* **294**, 4331–4344 (2019).

387. Singh, S. *et al.* SF3B1 mutations induce R-loop accumulation and DNA damage in MDS and

leukemia cells with therapeutic implications. *Leukemia* **34**, 2525–2530 (2020).

388. Barroso, S. The yeast and human FACT chromatin- reorganizing complexes solve R-loop-mediated transcription – replication conflicts. *Genes Dev.* **28**, 735–748 (2014).

389. Promonet, A. *et al.* Topoisomerase 1 prevents replication stress at R-loop-enriched transcription termination sites. *Nat. Commun.* **11**, 1–12 (2020).

390. Yoneyama-Hirozane, M. *et al.* High-Throughput Screening to Identify Inhibitors of DEAD Box Helicase DDX41. *SLAS Discov.* **22**, 1084–1092 (2017).

391. Nojima, T. *et al.* RNA Polymerase II Phosphorylated on CTD Serine 5 Interacts with the Spliceosome during Co-transcriptional Splicing. *Mol. Cell* **72**, 369-379.e4 (2018).

392. Muniz, L., Nicolas, E. & Trouche, D. RNA polymerase II speed: a key player in controlling and adapting transcriptome composition. *EMBO J.* **40**, 1–21 (2021).

393. Weissmann, S. *et al.* Landscape of TET2 mutations in acute myeloid leukemia. *Leukemia* **26**, 934–942 (2012).

394. Gaidzik, V. I. *et al.* TET2 mutations in acute myeloid leukemia (AML): results from a comprehensive genetic and clinical analysis of the AML study group. *J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol.* **30**, 1350–1357 (2012).

395. Ley, T. J. *et al.* DNMT3A mutations in acute myeloid leukemia. *N. Engl. J. Med.* **363**, 2424–2433 (2010).

396. Park, D. J. *et al.* Characteristics of DNMT3A mutations in acute myeloid leukemia. *Blood Res.* **55**, 17–26 (2020).

397. Chou, W. C. *et al.* TET2 mutation is an unfavorable prognostic factor in acute myeloid leukemia patients with intermediate-risk cytogenetics. *Blood* **118**, 3803–3810 (2011).

398. Richardson, S. E. & Huntly, B. J. P. Targeting Chromatin Regulation in Acute Myeloid Leukemia. *HemaSphere* **5**, (2021).

399. Qu, S. *et al.* Molecular and clinical features of myeloid neoplasms with somatic DDX41 mutations. *Br. J. Haematol.* 1–5 (2020) doi:10.1111/bjh.16668.

400. Tothova, Z. *et al.* Cohesin mutations alter DNA damage repair and chromatin structure and create therapeutic vulnerabilities in MDS/AML. *JCI Insight* **6**, (2021).

401. Smith, J. S. *et al.* Chronic loss of STAG2 leads to altered chromatin structure contributing to de-

regulated transcription in AML. *J. Transl. Med.* **18**, 1–18 (2020).

402. Cabal-Hierro, L. *et al.* Chromatin accessibility promotes hematopoietic and leukemia stem cell activity. *Nat. Commun.* **11**, 1–18 (2020).

403. Jalili, M. *et al.* Prognostic Value of RUNX1 Mutations in AML: A Meta-Analysis. *Asian Pac. J. Cancer Prev.* **19**, 325–329 (2018).

404. Gaidzik, V. I. *et al.* RUNX1 mutations in acute myeloid leukemia are associated with distinct clinico-pathologic and genetic features. *Leukemia* **30**, 2160–2168 (2016).

405. Geyh, S. *et al.* Transforming growth factor β 1-mediated functional inhibition of mesenchymal stromal cells in myelodysplastic syndromes and acute myeloid leukemia Correspondence : **103**, 1462–1471 (2018).

406. Liu, N., Zhang, J. & Ji, C. The emerging roles of Notch signaling in leukemia and stem cells. *Biomark. Res.* **1**, 1–7 (2013).

407. Tohda, S. NOTCH signaling roles in acute myeloid leukemia cell growth and interaction with other stemness-related signals. *Anticancer Res.* **34**, 6259–6264 (2014).

408. Jing Ma, Nadim Mahmud, Maarten C. Bosland, S. R. R. DDX41 is needed for pre-and post-natal hematopoietic stem cell differentiation in mice. (2021).

409. Chen, J.-Y. *et al.* Systematic Evaluation of Different R-loop Mapping Methods: Achieving Consensus, Resolving Discrepancies and Uncovering Distinct Types of RNA:DNA Hybrids. *bioRxiv* 2022.02.18.480986 (2022) doi:10.1101/2022.02.18.480986.

410. Bai, X., Li, F. & Zhang, Z. A hypothetical model of trans-acting R-loops-mediated promoter-enhancer interactions by Alu elements. *J. Genet. Genomics* **48**, 1007–1019 (2021).

411. Hartono, S. R. *et al.* The Affinity of the S9.6 Antibody for Double-Stranded RNAs Impacts the Accurate Mapping of R-Loops in Fission Yeast. *J. Mol. Biol.* **430**, 272–284 (2018).

412. Chen, Y.-H. *et al.* Transcription shapes DNA replication initiation and termination in human cells. *Nat. Struct. Mol. Biol.* **26**, 67–77 (2019).

413. Schärer, O. D. Nucleotide Excision Repair in Eukaryotes. *Cold Spring Harb. Perspect. Biol.* **5**, (2013).

414. Ong, S.-E. *et al.* Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics* **1**, 376–386 (2002).

415. Michalski, A. *et al.* Mass spectrometry-based proteomics using Q Exactive, a high-performance benchtop quadrupole Orbitrap mass spectrometer. *Mol Cell Proteomics* **10**, M111 011015 (2011).

416. Kelstrup, C. D., Young, C., Lavallee, R., Nielsen, M. L. & Olsen, J. V. Optimized fast and sensitive acquisition methods for shotgun proteomics on a quadrupole orbitrap mass spectrometer. *J Proteome Res* **11**, 3487–3497 (2012).

417. Olsen, J. V *et al.* Higher-energy C-trap dissociation for peptide modification analysis. *Nat Methods* **4**, 709–712 (2007).

418. Cox, J. *et al.* Andromeda: A peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–1805 (2011).

419. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **4**, 207–214 (2007).

420. Franceschini, A. *et al.* STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **41**, D808-15 (2013).

421. Saito, R. *et al.* A travel guide to Cytoscape plugins. *Nat Methods* **9**, 1069–1076 (2012).

422. Kuleshov, M. V. *et al.* Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90–W97 (2016).

423. Schindelin, J. *et al.* Fiji - an Open platform for biological image analysis. *Nat. Methods* **9**, (2009).

424. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).

425. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

426. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

427. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).

428. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12**, 115–121 (2015).

429. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160-5 (2016).

430. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

431. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, (2008).

432. Yu, G., Wang, L. G. & He, Q. Y. ChIP seeker: An R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics* **31**, 2382–2383 (2015).

433. Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinforma.* **47**, 11.12.1-34 (2014).

434. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).

435. Fridy, P. C. *et al.* A robust pipeline for rapid production of versatile nanobody repertoires. *Nat. Methods* **11**, 1253–1260 (2014).

436. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal; Vol 17, No 1 Next Gener. Seq. Data Anal. - 10.14806/ej.17.1.200* (2011).