UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Genome-wide Identification and Analysis of R-loop

Permalink

https://escholarship.org/uc/item/3wt8484f

Author

Zhang, Xuan

Publication Date

2021

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA SAN DIEGO

Genome-wide Identification and Analysis of R-loop

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Chemistry

by

Xuan Zhang

Committee in charge:

Professor Xiang-Dong Fu, Chair
Professor Gourisankar Ghosh, Co-Chair
Professor Michael G. Rosenfeld
Professor Susan S. Taylor
Professor Faik Akif Tezcan

2021

Copyright

Xuan Zhang, 2021

All rights reserved.

The dissertation of Xuan Zhang is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

To my parents.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

First, and foremost, I would like to thank my advisor, Professor Xiang-Dong Fu, for his support and guidance throughout my graduate career. Thanks for always listening to my ideas and offering encouragement and suggestions at crucial times. His brilliant and creative enthusiasm for science, superior intelligence and perseverance have inspired me and will keep motivating me in my future life.

I also sincerely thank Yajing Hao and Jun Xu for their support and guidence on many occasions in the past years. Fu lab members have all contributed to my education here at UCSD. A number of my friends at the department have helped with my project by being willing to test it and offer further ideas. Zhenze Jiang deserves special thanks in this respect.

Dongyang Zhang deserves my gratitude for listening to me during the trying moments of a Ph.D.. Of course any acknowledgment would be incomplete without expressing my deep gratitude to my parents for everything they have done for me.

Chapter 1, in part, are reprints of the material as it appears in X. Zhang, Y. Hao, X.-D. Fu,"Mapping R-loops using catalytically inactive RNaseH1 (R-ChIP)", *Methods in Molecular Biology*, 2021, The dissertation author was the primary investigator and author of this paper.

Chapter 2, in part, are reprints of the material as it appears in X. Zhang, Y. Hao, X.-D. Fu,"Mapping R-loops using catalytically inactive RNaseH1 (R-ChIP)", *Methods in Molecular Biology*, 2021, J.-Y. Chen, X. Zhang, X.-D. Fu, and L. Chen., "R-Chip for genome-wide mapping of r-loops by using catalytically inactive RNASEH1.", *Nature Protocols,*, 2019, L. Chen, J.-Y. Chen, X. Zhang, Y. Gu, R. Xiao, C. Shao, P. Tang, H. Qian, D. Luo, H. Li, Y. Zhou, D.-E. Zhang, and X.-D. Fu. "R-chip using inactive RNase H reveals dynamic coupling of r-loops with transcriptional pausing at gene promoters." *Molecular cell*, 2017, The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full is currently being prepared for submission for publication of the material. X. Zhang, J. Xu, Y. Hao, D. Wang, X.-D. Fu, "CSB Resolves R-loops by resuming stalled Pol II

at poly T tract". The dissertation author was the primary investigator and author of this paper.

VITA

2015	B. S. in Chemistry, Nanjing University, China
2017	M. S. in Chemistry, University of California San Diego
2021	Ph. D. in Chemistry, University of California San Diego

PUBLICATIONS

X. Zhang, J. Xu, Y. Hao, D. Wang, X.-D. Fu,"CSB Resolves R-loops by resuming stalled Pol II at poly T tract", In preparation.

X. Zhang, Y. Hao, X.-D. Fu,"Mapping R-loops using catalytically inactive RNaseH1 (R-ChIP)", *Methods in Molecular Biology*, In press.

J.-Y. Chen, X. Zhang, X.-D. Fu, and L. Chen., "R-chip for genome-wide mapping of r-loops by using catalytically inactive RNASEH1.", *Nature Protocols,*,14(5):1661, 2019

L. Chen, J.-Y. Chen, X. Zhang, Y. Gu, R. Xiao, C. Shao, P. Tang, H. Qian, D. Luo, H. Li, Y. Zhou, D.-E. Zhang, and X.-D. Fu. "R-chip using inactive rnase h reveals dynamic coupling of r-loops with transcriptional pausing at gene promoters." *Molecular cell*, 68(4):745–757, 2017

H. Qian, X. Kang, J. Hu, D. Zhang, Z. Liang, F. Meng, X. Zhang, Y. Xue, R. Maimon, S. F. Dowdy, N.K. Devaraj, Z.Zhou, W.C. Mobley, D.W. Cleveland, and X.-D. Fu. "Reversing a model of parkinson's disease with in situ converted nigral neurons." *Nature*, 582(7813):550–556, 2020

L.-T. Gou, D.-H. Lim, W. Ma, B.E. Aubol, Y. Hao, X. Wang, J. Zhao, Z. Liang, C. Shao, X. Zhang, F. Meng, H. Li, X. Zhang, R. Xu, D. Li, M.G. Rosenfeld, P.L. Mellon, J.A. Adams, M.-F. Liu, and X.-D. Fu. "Initiation of parental genome reprogramming in fertilized oocyte by splicing kinase srpk1-catalyzed protamine phosphorylation." *Cell*, 180(6):1212-1227, 2020

L. Chen, J.-Y. Chen, Y.-J. Huang, Y. Gu, J. Qiu, H. Qian, C. Shao, X. Zhang, J. Hu, H. Li, S. He, Y. Zhou, O. Abdel-Wahab, D.-E. Zhang, and X.-D. Fu. "The augmented r-loop is a unifying mechanism for myelodysplastic syndromes induced by high-risk splicing factor mutations." *Molecular cell*, 69(3):412–425, 2018

ABSTRACT OF THE DISSERTATION

Genome-wide Identification and Analysis of R-loop

by

Xuan Zhang

Doctor of Philosophy in Chemistry

University of California San Diego, 2021

Professor Xiang-Dong Fu, Chair
Professor Gourisankar Ghosh, Co-Chair

R-loop, a three-stranded RNA/DNA structure, has been linked to induced genome instability and regulated gene expression. To enable precision analysis of R-loops in vivo, we develop an RNase-H-based approach; this reveals predominant R-loop formation near gene promoters with strong G/C skew and propensity to form G-quadruplex in non-template DNA, corroborating with all biochemically established properties of R-loops. Transcription perturbation experiments further indicate that R-loop induction correlates to transcriptional pausing. Interestingly, we note that most mapped R-loops are each linked to a nearby free RNA end; by using a ribozyme to co-transcriptionally cleave nascent RNA, we demonstrate that such a free RNA

end coupled with a G/C-skewed sequence is necessary and sufficient to induce R-loop. These findings provide a topological solution for RNA invasion into duplex DNA and suggest an order for R-loop initiation and elongation in an opposite direction to that previously proposed.

Increasing evidence suggests that R-loops participates in the transcriptional coupled repair (TCR) pathway, which all point to the interaction between the TCR signaling factor Cockayne syndrome group B (CSB) and R-loops. However, how CSB sense and bind R-loops remains unclear. We use R-ChIP, a high resolution and high accuracy R-loop profiling method, to reveal the underlying mechanism. Not simply in TCR, CSB is a universal regulator of R-loops during transcription, where its association is frequently detected at pol II pausing loci especially at gene body. Depletion of CSB induces a new type of R-loop associated with the poly T tract on non-template strand. CSB resolves the pol II pausing induced R-loops through pushing pol II forward is further validated by the in vitro transcription assay. Taken together, our results for the first time shows R-loops are induced at poly T tract, uncovering a unifying mechanism of R-loop formation at pol II pausing sites. Failure to resolve such transient R-loops during transcription by CSB, would impact long genes transcription and potentially cause more DNA damages.

Chapter 1

Introduction

1.1 R-loop and its function

R-loops were first observed during experiments attempting to map DNA topology relative to expressed RNA. This technology was first applied to study the transcription unit for rRNA. By using purified rRNA to anneal to the corresponding DNA template, thereby looping out specific single-stranded DNA regions, it was possible to directly visualize the R-loop structure by electronic microscopy to deduce the gene structure [Thomas et al., 1976, White and Hogness, 1977]. This strategy played a vital role in the discovery of introns in a protein-coding gene from an adenovirus genome [Berget et al., 1977, Chow et al., 1977]. It was soon recognized that such structures are also formed on chromatin during important biological processes, such as immunoglobulin class switch and transcription [Roy et al., 2008, Belotserkovskii et al., 2018]. Further studies demonstrated that the formation of R-loops provides an important mechanism to

regulate gene expression during cell fate determination [Niehrs and Luke, 2020].

R-loop is a three-stranded nucleic acid structure made by a DNA/RNA hybrid and a looped-out DNA strand. Such DNA/RNA hybrid with a displaced ssDNA quite often exists within lots of enzymes like Cas9 or RNA pol II which need to bind a DNA/RNA hybrid. However, in most cases we call it as R-loop to indicate the longer RNA/DNA hybrid which is generated from an RNA molecule annealed to the antisense DNA strand after is exits the active site of the RNA polymerase. At the same time a displaced sense ssDNA is generated, known as the 'thread-back' model. Figure 1.1 [Thomas et al., 1976] shows the R-loop under the electron microscope by mixing dsDNA with RNA in the presence of formamide.

R-Loops are considered as byproducts which threats to genome stability because of the exposed ssDNA. (Figure 1.2) Once the R-loop is produced, it can be removed by degradation of the RNA strand. This is achieved by RNASE H1, which specifically degrades the RNA moiety within RNA-DNA hybrids [Nowotny et al., 2007]. Alternatively, R-loops can be removed by RNA-DNA helicase such as the recG which is a DNA helicase in E. Coli [Hong et al., 1995, Harinarayanan and Gowrishankar, 2003]. The yeast protein Sen1 and its human homologue senataxin were initially identified as a DNA and RNA helicase with 5' to 3' RNA-DNA unwinding activity in vitro, which has been implicated in R-loop homeostasis [Kim et al., 1999].

There are also some mechanisms to prevent the formation of R-loop. Top1 can resolve local negative supercoiling behind the elongating RNA Pol II. The negative supercoiling may lead to a transient local unwinding of the DNA strand.

Although studies in the last decade were concentrated on the detrimental effects of R-loop formation, particularly on genome stability. But recent data have revealed that R-loops can also have a positive impact on cell processes, like regulating gene expression, chromosome structure

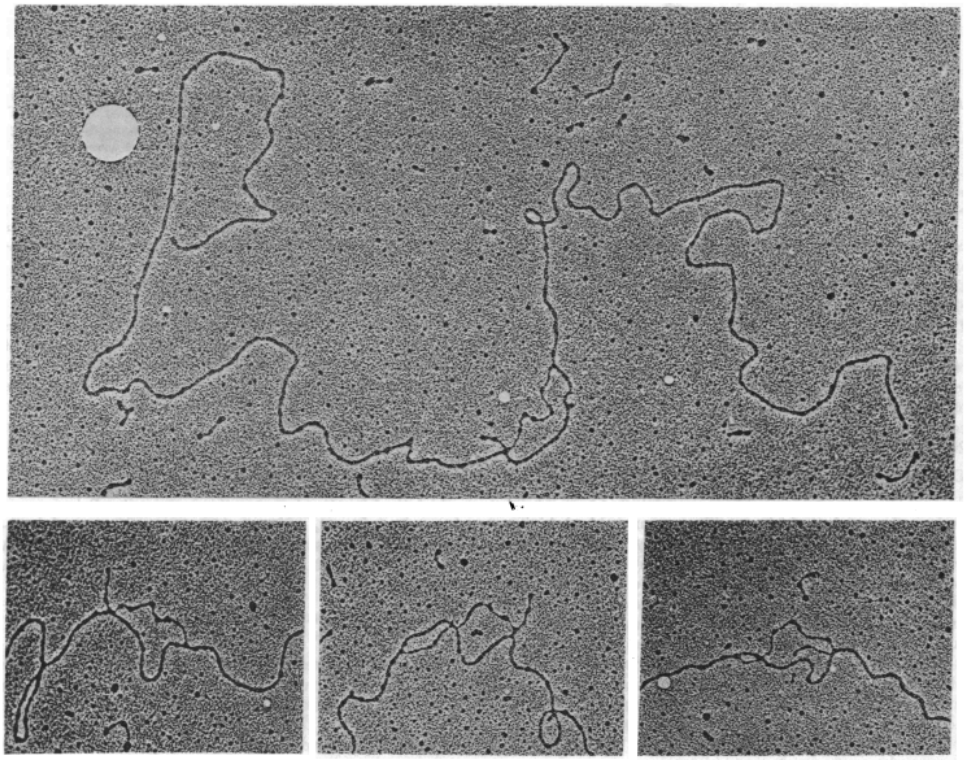

Figure 1.1: **R-loop visulization under electron microscope.** R-loops were made by heating 5 μg/ml of λgt-Sc11O9 DNA and 5 μg/ml total rRNA in 70% vol/vol formamide, 0.1 M Pipes at pH 7.8, and 0.01 M Na_3EDTA at 47°for 20 hr. The reaction was performed under oil in a sealed, siliconized glass tube. All 500 RNA molecules examined contained an R-loop similar to those shown. The sample was mounted for electron microscopy by the formamide technique (2). Grids were stained with uranyl acetate and shadowed with Pt/Pd.This figure is from [Thomas et al., 1976]

.

and DNA repair. In Arabidopsis, it has been indicated the formation of R-loops in the promoter region silences the expression of the long noncoding RNA COOLAIR which in turns regulates the flowering locus in response to cold temperatures[Hawkes et al., 2016].

1.2 R-loop identification

To understand where such R-loops are induced in the genome, the field has largely relied on the use of a monoclonal antibody (S9.6) that shows high specificity for DNA:RNA hy-

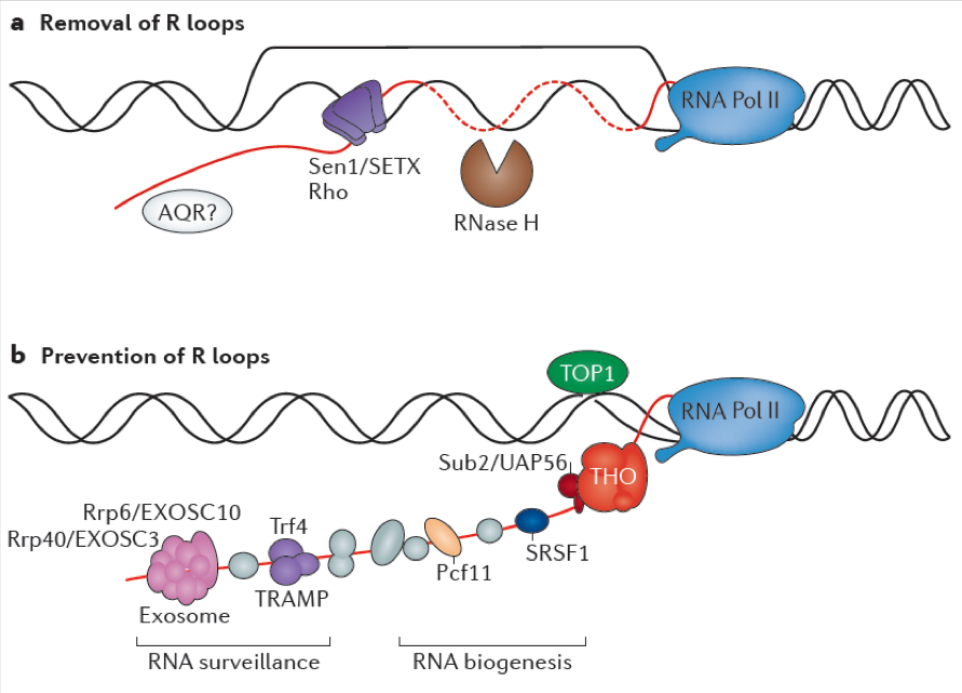

Figure 1.2: R-loop dynamics regulation. This figure is from [Santos-Pereira and Aguilera, 2015]

.

brids [Chédin et al., 2021]. This antibody has been employed to immunoprecipitate R-loops containing DNA or RNA coupled with deep sequencing to map potential R-loops genome-wide, technologies known as DRIP-seq or DRIPc-seq. However, we and others have been puzzled by inconsistent data from different labs, many of which appear to contradict with one another and to biochemically characterized R-loops. Realizing a series of potential artifact-prone sources with the existing technologies, we went on to develop our own strategy to map R-loops genome-wide by using a catalytically inactive RNase H, an enzyme evolutionarily evolved to specifically recognize DNA-RNA hybrids in the cell. This new technology, called R-ChIP[Chen et al., 2017], simply expresses a catalytically dead RNase H1 gene in cells followed by using standard ChIP-seq to map its binding sites in the genome, revealing predominant R-loop formation at active gene promoters.

Importantly, by using the new technology to fine map R-loops in mammalian genomes, we have revised a long-hold model for R-loop initiation and progression. A key finding is a prerequisite for a free 5' end for nascent RNA to invade into DNA duplex in GC-skewed sequence context. We discovered that R-loop formation is directly linked to transcriptional pausing at gene promoters. Interestingly, we further found that, contrary to a tight link between R-loop induction and transcription pausing, R-loop resolution is neither necessary nor sufficient for transcription pause release, suggesting that R-loop resolution and transcriptional pause release are separately regulated processes.

1.3 R-loop and disease

1.3.1 R-loop and MDS

Several key splicing factors (SRSF2, U2AF1, etc.) have recently emerged as major disease genes in leukemia and solid tumors [Dvinge et al., 2016]. Each causal mutation induces widespread non-overlapping changes in alternative splicing [Qiu et al., 2016, Komeno et al., 2015]. This new technology we developed for mapping R-loops genome-wide reveals all high-risk alleles in splicing factors cause excessive R-loop formation, which activates the ATR/Chk1 pathway to induce DNA damage response and cell cycle arrest [Chen et al., 2018].

MDS is a highly heterogeneous disease that causes dysplasia of hematopoietic progenitor cells, implying that multiple mechanisms may underlie the disease etiology and/or progression. A major advance in recent years is the identification of prevalent mutations in several "general" splicing factors, leading to research focusing on altered splicing induced by individual splicing

factor mutations as potential causes for the disease. Indeed, it has been shown that the inclusion of a toxic exon in the Ezh2 transcript in SRSF2(P95H) knockin mice causes downregulation of the EZH2 protein via nonsense-mediated mRNA decay, and overexpression of EZH2 was able to restore the proliferative potential of hematopoietic progenitors [Kim et al., 2015]. However, each splicing factor mutation appears to affect a largely distinct set of splicing, as we now further confirmed in the same cellular background, raising the question of whether and how multiple independent splicing changes all cause a similar disease phenotype.

We found that mutations in splicing factors may contribute to MDS via induced R-loop formation [Chen et al., 2018]. Importantly, mutations in SRSF2 and U2AF35 induce R-loops at gene promoters, which is functionally linked to defects in transcriptional pause release.

1.3.2 R-loop and Cockayne Syndrome

Accurate maintenance of the whole genome is vital for the cell function, especially for active genes. However, the frequent and complex transcription happened at active genes, may induce more DNA damages [Kim and Jinks-Robertson, 2012]. TCR guarantees the genome stability as it initiates DNA repair pathways when lesion is detected during transcription [Hanawalt and Spivak, 2008]. R-loops are frequently formed in such transient and unscheduled co-transcriptional processes. Not simply be found as by-products, R-loops also function as important regulators at DNA damaged sites induced by various treatments. For example, Rad52 is recruited to ionizing radiation induced DSBs in a DNA/RNA hybrid-dependent manner, which then initiates subsequent repair by homologous recombination [Yasuhara et al., 2018]; RBM14 binds directly to the DNA/RNA hybrid at the endonuclease induced DSBs, which is required for the canonical non-homologous end joining (cNHEJ) repair [Jang et al., 2020]. However, the mechanism of how

R-loops accumulated around DSB remains unknown.

In TCR, CSB is one of the first proteins to be recruited to the arrested Pol II, sending signals to initiate the downstream DNA repair pathways [Xu et al., 2017a]. Mutations in CSB cause Cockayne syndrome (CS), an autosomal-recessive neurological disorder with the phenotype of premature aging, photosensitivity and growth failure, all linking with genome instability, which are due to th fact that CSB deficient cells are unable to initiate TCR [Karikkineth et al., 2017]. Besides, CSB also plays important roles in transcription regulation by exerting function in chromatin remodeling and transcription pausing release[Xu et al., 2020, Xu et al., 2017a]. As expected, it has been found that CSB interacted with R-loops in multiple TCR pathways. For example, in the TC-HR activated by the ROS induced DSBs at transcribed loci, CSB is recruited by R-loops first, providing the interface to further recruit downstream repair factor, Rad52 [Teng et al., 2018].

We demonstrate that R-loops are significantly induced at transcription start site and gene body when CSB is depleted. Moreover, a new type of R-loops associated with poly T tract on the non-template strand are very sensitive to CSB depletion, indicating that CSB are recruited to the poly T loci during transcription. Importantly, GRO-seq signal decreased sharply at the poly T tract, also confirmed by in vitro transcription assay, showing that such poly T associated R-loops are induced due to pol II pausing. These findings suggest that CSB not only function as a TRC initiation factor, but also a global R-loop regulator. It also provides new insights into the underlying mechanism of the Cockayne syndrome.

1.4 Acknowledgements

Chapter 1, in part, are reprints of the material as it appears in X. Zhang, Y. Hao, X.-D. Fu,“Mapping R-loops using catalytically inactive RNaseH1 (R-ChIP)”, *Methods in Molecular Biology*, 2021, The dissertation author was the primary investigator and author of this paper.

Chapter 2

Mapping R-loops using inactive RNase H1

2.1 Introduction

Mapping R-loop locations along the genome is key to understand many critical biological processes. In the past decade, multiple methods have been developed to map R-loops genome-wide. Most of them have been largely based on the use of a monoclonal antibody (S9.6) specific for RNA/DNA hybrids. In recent years, this affinity probe has been coupled with deep sequencing to detect R-loops genome-wide in yeast, plant, and mammals [Niehrs and Luke, 2020]. The majority of these methods use the S9.6 antibody to capture RNA/DNA hybrids from restriction digested genomic DNA followed by deep sequencing of captured DNA (DNA:RNA immunoprecipitation [DRIP] sequencing [DRIP-seq] and its derivatives) or RNA (DNA:RNA immunoprecipitation followed by cDNA conversion [DRIPc]-seq and its derivatives) [Chédin et al., 2021]. Because of limited resolution with restriction digestion, additional efforts have also been made

to use sonication to increase R-loop mapping resolution; however, such treatment may destroy some fragile R-loops in the absence of fixation [Chen et al., 2019]. These technical problems may thus account for various inconsistent results in the literature. For example, a published R-loop profile showed a metagene peak near gene promoters [Chen et al., 2015a], but other studies suggested R-loops are more spread in gene bodies [Ginno et al., 2012], and yet another work revealed a metagene peak 1.5 Kb downstream of transcription start sites (TSSs) [Sanz et al., 2016]. While most mapped R-loops are linked to G/C skew, a recent report even detected A/T skew underlying mapped R-loops in yeast [Wahba et al., 2016].

Because location matters for understanding R-loop biology, we sought to develop an in vivo strategy for R-loop profiling based on RNase H, a gold standard for R-loop recognition in the cell. By expressing a catalytically dead *RNaseH1* followed by strand-specific amplification of immunoprecipitated (IPed) DNA (termed R-ChIP), we efficiently captured R-loops associated with all key sequence features established by biochemistry. Transcriptional perturbation experiments revealed a dynamic link of R-loop induction with RNAPII pausing at gene promoters. We also noted the proximity of most mapped R-loops to free RNA ends. We pursued this intriguing association by using an engineered ribozyme system to show the functional requirement of a free RNA end coupled with a G/C-skewed sequence for promoting de novo R-loop formation. These findings have important implications in the requirement for R-loop initiation and elongation as well as the contribution of R-loops to transcriptional control.

2.2 In vivo R-loop profiling by inactive RNaseH1

Current R-loop profiling methods appear to have an intrinsic limitation in pinpointing the exact location of R-loops in individual genomic fragments captured by S9.6 according to a recent analysis [Halász et al., 2017]. Diminished signals upon RNase H digestion may reflect the absence of R-loop for capture but do not necessarily show where a specific R-loop(s) is in individual restriction fragments. Although RNase H has been previously explored to capture RNA/DNA hybrids in vitro (DRIVE-seq), the capture efficiency appears quite poor compared to S9.6. Because RNase H may recognize R-loops more efficiently and with higher specificity in the cell than in test tube and a catalytically dead RNase H has been demonstrated to target R-loops in vivo, we were motivated to explore this in vivo R-loop capture strategy to understand the mechanism for R-loop formation.

Mammalian cells express two RNase H enzymes, one of which (RNASEH1) is composed of a single polypeptide, thus convenient for molecular manipulation. Because RNase H is known to target both the nucleus and mitochondrion, we first constructed two forms of human RNASEH1 mutant proteins fused with a nuclear localization signal (NLS) at the N terminus for exclusive targeting to the nuclear genome and a V5 tag at the C terminus to enable efficient IP (Figure 2.1). The D210N mutation abolishes the catalytic activity of RNASEH1, whereas a combination of three specific mutations (W43A, K59A, K60A) in the binding domain prevents the enzyme from binding to RNA/DNA hybrids. A quadruple mutant that carries all four mutations (WKKD) thus inactivates both binding and catalytic activities of RNASEH1, which provides an ideal negative control. We introduced both wild-type (WT) and mutant RNASEH1 (D210N and WKKD) into HEK293T cells, all expressed at comparable levels. Immunocytochemistry showed

exclusive nuclear localization of all tagged RNASEH1 proteins with selective enrichment in the nucleolus (Figure 2.2), consistent with extensive R-loops on rDNA loci. Despite overexpression, neither WT nor mutant RNASEH1 caused any measurable defects in cell proliferation and cell cycle progression, in line with previous reports. In theory, a catalytically dead RNASEH1 may bind to R-loop and prevent its resolution, but we did not detect obvious impact on nascent RNA production measured by global run-on sequencing (GRO-seq) when comparing cells with or without expressing the D210N mutant.

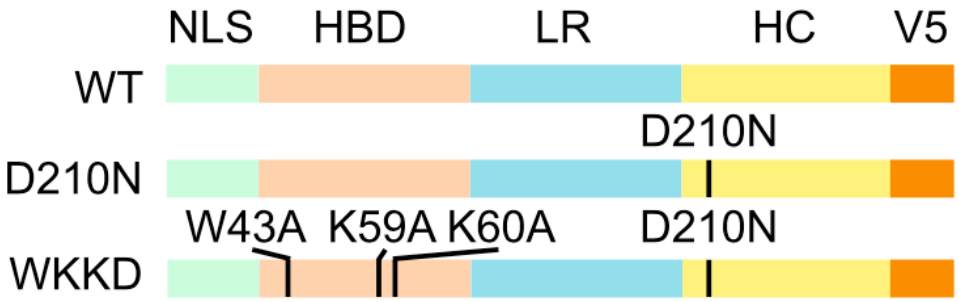

Figure 2.1: Design of RNASEH1 expression vectors. NLS, nuclear localization signal; HBD, RNA/DNA hybrid binding domain; LR, linker region; HC, RNA/DNA hybrid catalytic domain; V5, V5 tag.

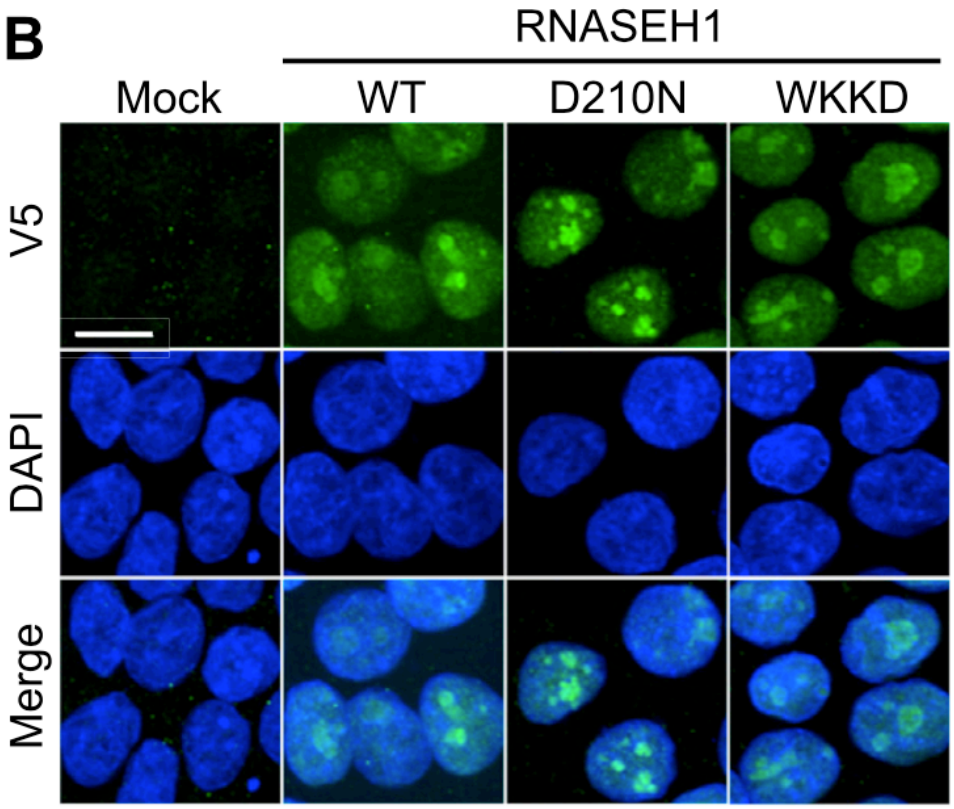

Figure 2.2: Localization of exogenously expressed wild-type (WT) and mutant (D210N and WKKD) RNASEH1 in HEK293T cells by immunocytochemistry. Green: V5, Blue: DAPI; scale bar, 20 mm.

Next, we designed a strand-specific strategy to analyze captured R-loops in vivo. As illustrated in Figure 2.3, we first performed chromatin immunoprecipitation (ChIP) with an anti-V5 antibody. The IPed DNA part of the RNA/DNA hybrid was next converted to double-strand DNA (dsDNA) by using an adaptor containing random primer followed by ligation to another double-stranded adaptor for PCR amplification and deep sequencing. Although both D210N and WKKD mutant proteins were IPed with equal efficiency with anti-V5 (Figure 2.4, 2.5), ChIP-qPCR analysis revealed strong signals at TSS regions of multiple genes in cells expressing D210N but not WKKD, indicating that the catalytically dead RNASEH1 can specifically target R-loops in vivo. We thus proceeded with deep sequencing of R-ChIP libraries, yielding a total of 49 million 40-nt

reads (22.6 million uniquely mapped reads after removing PCR duplicates) from three biological replicates, and demonstrated high global reproducibility by pairwise comparison of independent R-ChIP libraries.

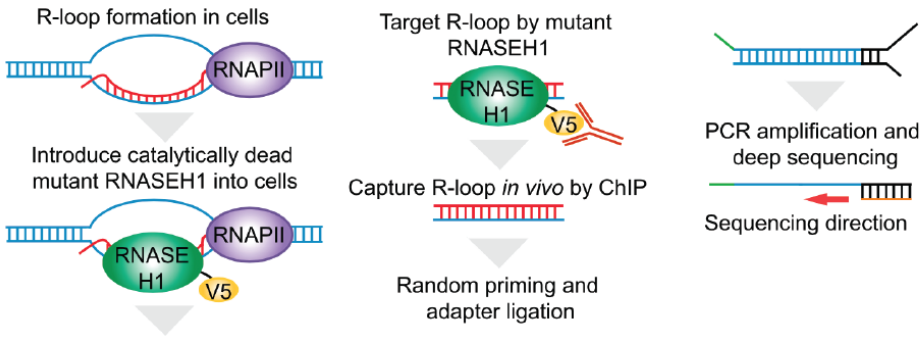

Figure 2.3: Schematic presentation of the R-ChIP strategy.

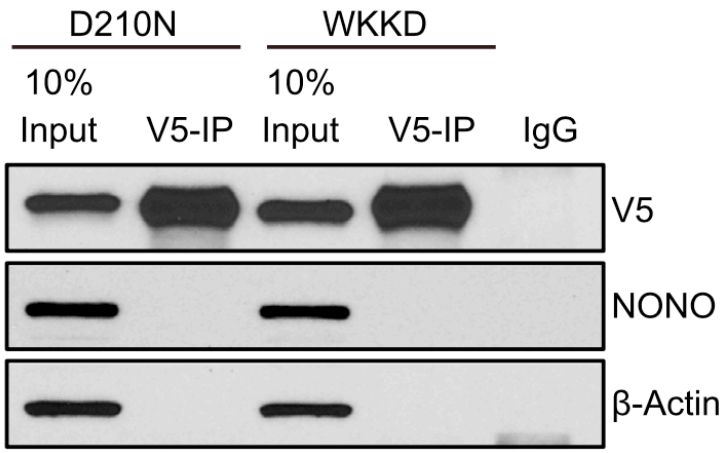

Figure 2.4: Immunoprecipitation of exogenously expressed RNASEH1 (D210N and WKKD mutants) by using anti-V5 antibody with similar efficiency. The levels of RNASEH1 were analyzed by western blotting relative to invariant nuclear protein NONO and cytoplasmic protein b-actin.

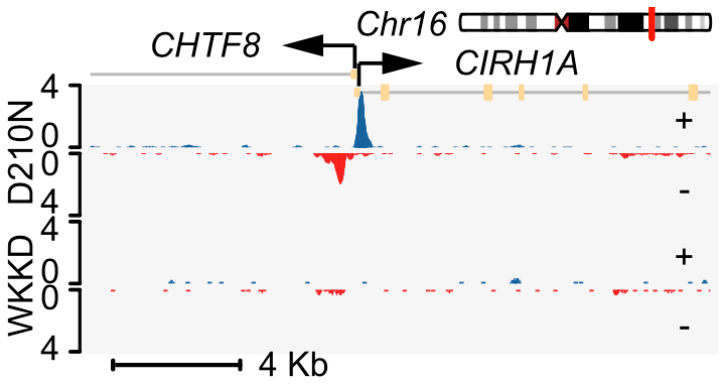

Figure 2.5: A representative genomic region showing the R-ChIP signals in cells expressing D210N or WKKD mutant proteins. Blue: + strand, Red: – strand.

Using a standard ChIP-seq peak calling method, we identified 12,906 specific peaks based on total input control. As expected, we detected specifically enriched signals with D210N, but not WKKD, on a specific example as well as globally (Figure 2.6). Because of some low levels of ChIP-seq signals with WKKD, we subtracted such background and found that out of 12,906 R-ChIP peaks, 12,521 (97%) still remained as significant peaks. The mapped R-loops were in line with the orientation of locally transcribed RNAs detected by GRO-seq (Figure 2.6) and showed remarkable strand specificity (Figure 2.7). To further validate specific D210N binding events in the genome, we also engineered a HEK293T cell line expressing a truncated form of RNASEH1 with the catalytic domain deleted and found that this DHC mutant generated specific R-ChIP signals essentially identical to those detected with D210N. Together, these data strongly suggest that R-ChIP efficiently and faithfully captures R-loops in vivo.

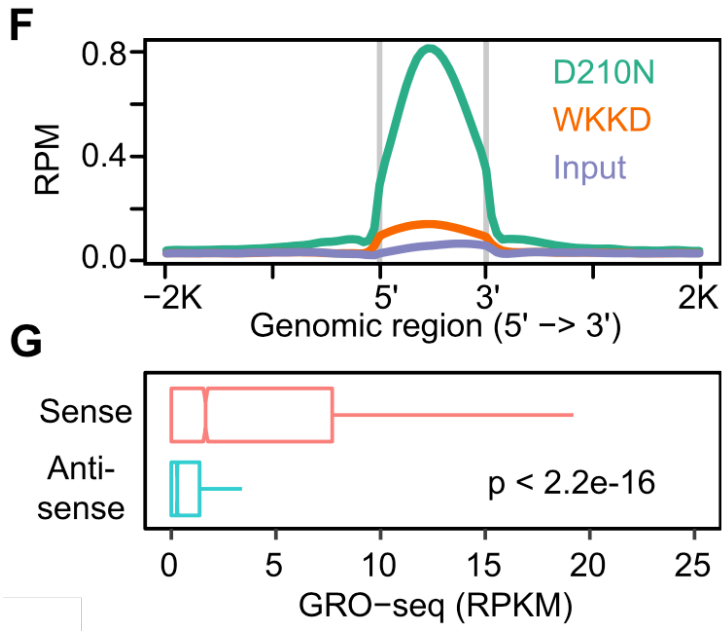

Figure 2.6: The signal intensity profiles of R-ChIP within the peak regions in HEK293T cells expressing D210N versus WKKD relative to input. The signal intensity of GRO-seq from the same (sense) or opposite (anti-sense) strand of individual R-ChIP peak regions. Wilcoxon test was used to calculate the p value.

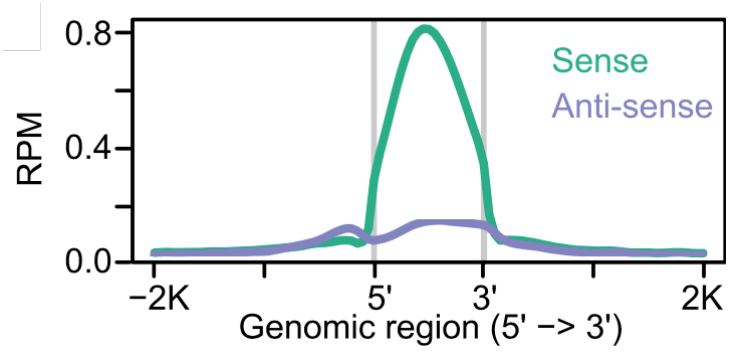

Figure 2.7: The strand specificity of R-ChIP signals. The signals associated with a R-ChIP peak were divided into sense (in the same orientation of the peak) and antisense (in the opposite orientation of the peak) groups for comparison.

2.3 R-ChIP Protocol

2.3.1 Materials

1. Crosslinking and sonication

    1.1 1x PBS

    1.2 37% Formaldehyde

    1.3 1.375 M Glycine

    1.4 RNaseOut (Invitrogen)

    1.5 Cell Lysis Buffer: 10 mM Tris-HCl pH 8.0, 10 mM NaCl, 0.5% NP-40, Protease Inhibitor Cocktail (Sigma, added immediately prior to use)

    1.6 Nuclear Lysis Buffer: 50 mM Tris-HCl pH 8.0, 1% SDS, 10mM EDTA, Protease Inhibitor Cocktail (added immediately prior to use)

    1.7 TE Buffer: 10 mM Tris-HCl pH 8.0, 1 mM EDTA

2. Chromatin Immunoprecipitation

    2.1 Dilution Buffer: 20 mM Tris-HCl pH 8.0, 150 mM NaCl, 2 mM EDTA, 1% Triton X-100

    2.2 Blocking Buffer: 10 mg/ml Glycogen, 10 mg/ml BSA, 20 mg/ml tRNA

    2.3 Magnetic beads (Dynabeads Protein G, Invitrogen)

    2.4 Antibody against the V5 tag (Invitrogen, cat. no. R960-25)

2.5 TSEI Buffer: 20 mM Tris-HCl pH 8.0, 150 mM NaCl, 2 mM EDTA, 1% Triton X-100, 0.1% SDS, Protease Inhibitor Cocktail (added immediately prior to use)

2.6 TSEII Buffer: 20 mM Tris-HCl pH 8.0, 500 mM NaCl, 2mM EDTA, 1% Triton X-100, 0.1% SDS, Protease Inhibitor Cocktail (added immediately prior to use)

2.7 Buffer III: 10 mM Tris-HCl pH 8.0, 250 mM LiCl, 1 mM EDTA, 1% NP-40, 1% sodium deoxycholate, Protease Inhibitor Cocktail (added immediately prior to use)

2.8 Elution Buffer: 10 mM Tris-HCl pH 8.0, 1 mM EDTA, 1% SDS

2.9 RNase A

2.10 Proteinase K

2.11 Phenol/chloroform/isoamyl alcohol (25:24:1)

3. Library construction and sequencing

3.1 dNTP mix

3.2 Phi29 DNA polymerase (NEB)

3.3 dATP

3.4 Klenow fragment (3′->5′exo-) (NEB)

3.5 T4 DNA Ligase and T4 DNA Ligase Buffer (NEB)

3.6 N9 random primer (Illumina HiSeq 2500 system):

5′-/Invddt/CAAGCAGAAGACGGCATACGAGNNNNNNNNN-3′

3.7 Adapter oligos (Illumina HiSeq 2500 system)

Oligo A: 5′-/Phos/GATCGGAAGAGCGTCGTGTAGGGAAAGAGTGT-3′

Oligo B: 5′-AGACGTGTGCTCTTCCGATCT-3′

3.8 Annealing Buffer: 10 mM Tris-HCl pH 8.0, 50 mM NaCl, 1 mM EDTA

3.9 PCR Primer Mix (Illumina HiSeq 2500 system)

PCR primer: 5′-CAAGCAGAAGACGGCATACGAG-3′

Barcode primer:

5′-AATGATACGGCGACCACCGAGATCTACACNNNNN

ACAC TCTTTCCCTACACGACGCTCTTCCGATCT-3′

3.10 Phusion DNA Polymerase (NEB)

3.11 DNA Cleanup Kit (Zymo)

3.12 10% TBE Gel

3.13 Gel loading buffer (NEB)

3.14 SYBR gold dye (Invitrogen)

3.15 GlycoBlue (Invitrogen)

3.16 Gel elution buffer: 10 mM Tris-HCl pH 8.0, 300 mM NaCl, 1 mM EDTA, 0.1% Tween 20

4. General supplies and equipment

4.1 Cell scraper (if working with adherent cells)

4.2 0.22 μm syringe filters

4.3 Probe sonicator or equivalent (See Notes 1)

4.4 Magnetic stand

4.5 1.5 mL microcentrifuge tubes, preferably low protein and DNA binding

4.6 Qubit fluorometer or equivalent (Thermo Fisher)

4.7 Qubit dsDNA HS Assay Kit (Invitrogen)

4.8 Costar column (Corning)

2.3.2 Methods

R-ChIP requires the construction of stable cell line expressing inactive RNaseH1, which can be delivered via either lentiviral virus or transposon. We used the pPyCAG vector to construct stable cell lines on HEK293T and K562 cells. Western blot should be used to confirm the inactive RNaseH1 expression with V5 antibody before immunoprecipitation. Buffers should be freshly prepared and kept on ice before use.

1. Cross-linking and sonication

Preparation of antibody conjugated magnetic beads (Day 1)

1.1 Add 50 μl of well-resuspended magnetic beads to 1.5 ml tube.

1.2 Wash the beads twice in ChIP Dilution Buffer. For each wash, resuspend the beads in 1ml ChIP Dilution Buffer, spin the tube briefly to collect cap and side wall liquid, incubate on the magnetic stand for 5 min, remove the supernatant.

1.3 Resuspend the beads in 1ml Blocking Buffer and incubate for 2 hours on a tube rotator at room temperature (RT).

1.4 Place the beads on a magnetic stand for 5 min and remove the supernatant.

1.5 Wash the beads three times with Blocking Buffer as described above. Resuspend beads in 1ml Blocking Buffer.

1.6 Add 5-10 μg of V5 antibody to the beads.

1.7 Incubate overnight with rotation at 4 °C.

Crosslinking

For each sample, $1\sim2*10^7$ *cells are enough for robust R-loop signal in K562 and HEK293T cells. Volumes below are described for one sample.*

1.8 Wash cells 3 times with ice-cold PBS, 10 ml PBS for each wash.

1.9 Add 270 μl 37% formaldehyde in 10 ml PBS to achieve final concentration at 1% (v/v). Mix well with cells and incubate for 10 min at RT.

1.10 Add 1 ml 1.375 M glycine to achieve final concentration at 0.125 M to quench the reaction. Mix well with cells and incubate for 15 min at RT.

1.11 Wash cells 3 times with ice-cold PBS, 10 ml PBS for each wash.

1.12 For adhesive cells, scrape the cells off the plate.

1.13 Collect cells by centrifugation at 600g at 4 °C for 5 min in 1.5 ml microcentrifuge tube. Discard the supernatant.

1.14 The half-dry cell pellet can be immediately frozen in liquid nitrogen and stored at -80 °C. The fixed R-loop stored at -80 °C is stable within 1 month.

Sonication

1.15 Add 1 ml ice-cold Cell Lysis Buffer (with protease inhibitor cocktail and RNaseOut added) to resuspend the frozen or fresh cell pellet in a 1.5 ml microcentrifuge tube. Incubate on ice for 10-20 min to release the nuclei.

1.16 Pellet the nuclei by centrifugation at 800g at 4 °C for 5 min. Discard the supernatant.

1.17 Wash the nuclei once with 1 ml Cell Lysis Buffer (with protease inhibitor cocktail and RNaseOut added) and pellet again as in Step 1.16.

1.18 Resuspend the nuclei in 500 μl Nuclear Lysis Buffer (with protease inhibitor cocktail and RNaseOut added). Mix well and incubate on ice for 10 min.

1.19 Sonicate in ice bath, by immersing the probe sonicator tip in the tube, using 7 cycles of 10 s sonication, 30 s rest between cycles, and power level of 4.0. Adjust the depth of the sonicator tip to avoid foaming. Check the temperature between cycles to make sure that the samples are not no overheated. The desired chromatin fragment size is between 200-500 bp before immunoprecipitation, which can be confirmed on 1.5% agarose gel.

1.20 Bury the sample tube in ice for 30-60 min to precipitate out the extra SDS from fragmented chromatin.

1.21 Centrifuge at 16000g at 4 °C for 15 min to pellet the cell debris and extra crystalized SDS.

1.22 Collect the supernatant and adjust the sample volume to 1 ml with TE buffer.

1.23 Add following additional reagents to bring the volume to 1.3 ml for each sample.

Stock Solution	Concentration	Volume/tube
10% Triton X-100	1%	130 μl
10% Sodium deoxycholate	0.1%	13 μl
100X Protease Inhibitor	1X	13 μl
1X TE	1X	144 μl

1.24 Save 50 μl "input protein" for western blot validation of IP efficiency, and 50 μl "input DNA" to provide input control for sequencing.

2. Immunoprecipitation

2.1 Place the antibody-conjugated beads prepared in Step 1.7 on the magnetic stand for 5 min and remove the supernatant.

2.2 Add 1.2 ml diluted chromatin from Step 1.24 to the beads. Incubate overnight with rotation at 4 °C.

Beads washing and decrosslinking (Day 2)

For each wash, resuspend the beads in 1 ml buffer, pipette the beads solution 20 times gently to mix well, place the beads solution on magnetic stand for 5 min, remove the supernatant, specifically,

2.3 Place the beads on a magnetic stand for 5 min. Collect the supernatant as the depleted R-loop solution for IP efficiency validation by western blot.

2.4 Wash the beads 3 times with TSEI Buffer.

2.5 Wash the beads 3 times with TSEII Buffer.

2.6 Wash the beads once with Buffer III.

2.7 Wash the beads once with 1 ml TE Buffer.

2.8 Centrifuge the beads at 1500g for 1 min. Remove the residual TE Buffer and collect the semi-dry beads.

2.9 Resuspend the beads in 170 μl Elution Buffer.

2.10 Incubate the beads solution at 65 °C for 30 min on the thermomixer to dissociate antibodies from the beads, vortex at 1200 rpm for 15 s with 2 min interval.

Decrosslinking

2.11 Place the beads on the magnetic stand for 5 min to collect the supernatant as immuno-precipitated R-loop sample.

2.12 Centrifuge at 16000g for 5 min at RT. Combine the residual supernatant to the IPed R-loop solution in the previous step. Aliquot 20 μl of the solution as the IPed protein for western blot analysis, together with the input protein collected at Step 1.24.

2.13 Add 120 μl Elution Buffer to the 50 μl "input DNA" prepared in Step 1.24.

2.14 Incubate both IPed R-loop solution and Input DNA at 65 °C overnight in a thermomixer, vortex at 1200 rpm for 15 s with 2 min interval.

DNA clean up (day 3)

2.15 Add 150 μl TE Buffer and 6 μl RNase A (10 mg/ml) to IPed R-loop and input samples.

2.16 Mix well and incubate for 2 hours at 37 °C in a thermomixer, vortex at 1200 rpm for 15 s with 2 min interval.

2.17 Add 7 μl of 20 mg/ml Proteinase K to IPed R-loop and input samples.

2.18 Mix well and incubate for 2 hours at 65 °C in a thermomixer, vortex at 1200 rpm for 15 s with 2 min interval.

2.19 Extract the DNA with phenol/chloroform/isoamyl alcohol twice in both IPed and input samples. For each extraction, add 300 μl phenol/chloroform/isoamyl alcohol, vortex for 20 s and centrifuge for 5 min at 16000g at RT, transfer the top (aqueous) phase to a fresh 1.5 ml DNA low-binding tube.

2.20 Add 750 μl ethanol, 30 μl of 3M NaAc and 1 μl GlycoBlue to IPed and input samples. Precipitate the DNA overnight at -20 °C or at -80 °C for 15 min.

2.21 Centrifuge at 16000g for 15 min at 4 °C to pellet the DNA.

2.22 Wash the pellet with 1 ml ice-cold 70% ethanol twice.

2.23 Dissolve the DNA pellet in 20 μl water.

2.24 Measure DNA concentration using the Qubit dsDNA HS assay kit.

3. Sequencing library construction

dsDNA Extension (day 4)

dsDNA extension converts the ssDNA generated from RNase A digested DNA/RNA hybrids to dsDNA by one cycle extension using N9 primer, specifically,

3.1 Prepare dsDNA extension reaction without enzyme in 200 μl PCR tube for each sample (annealing):

Reagent	Volume (19 μl)	Concentration
10 ng DNA (Step 2.23) diluted with water to 11 μl	11 μl	10 ng
10X phi29 DNA polymerase buffer	2 μl	1X
BSA (1 mg/ml)	4 μl	200 μg/ml
3 mM dNTP mix	1 μl	150 μM
20 μM N9 primer	1 μl	1 μM

3.2 Centrifuge briefly and incubate for 5 min at 95 °C, followed by 5 min at 25 °C in a thermal cycler.

3.3 Add 1 μl phi29 DNA polymerase and incubate for 20 min at 30 °C, followed by 10 min at 65 °C to inactive the enzyme.

3.4 Purify DNA using the Zymo DNA Cleanup Kit. Elute in 20 μl water.

dA Addition

3.5 Prepare dA addition reaction in a 200 μl PCR tube for each sample:

Reagent	Volume (30 μl)
DNA sample from Step 3.4	20 μl
10X Klenow buffer	3 μl
Klenow exo (3′ to 5′ exo minus)	1 μl
1 mM dATP	5 μl

3.6 Centrifuge briefly and incubate at 37 °C for 30 min in a thermal cycler.

3.7 Purify DNA using the Zymo DNA Cleanup Kit. Elute in 13 μl water.

Adapter Ligation

3.8 Adapter need to be pre-annealed before use. Resuspend adapter oligo A and B at a concentration of 100 μM in Annealing Buffer, respectively.

3.9 Mix 20 μl of 100 μM oligo A with 20 μl of 100 μM oligo B.

3.10 Add 60 μl Annealing Buffer.

3.11 Anneal oligos on a thermal cycler with the following program: 2 min at 95 °C, ramping down at the rate of 0.1 °C/s from 95 to 25 °C. The annealed adapters stock solution (20 μM) can be stored in -80 °C freezer.

3.12 Prepare 2 μM adapter solution by diluting 20 μM adapter stock solution 10 times with water.

3.13 Prepare Adapter Ligation reaction in a 200 μl PCR tube for each sample:

Reagent	**Volume (20 μl)**
DNA sample from Step 3.7	13 μl
10X ligation buffer	2 μl
Adapter A+B (2 μM)	1 μl
T4 DNA ligase	4 μl

3.14 Centrifuge briefly and incubate at RT for 2 hours.

PCR amplification

3.15 Prepare PCR reaction in a 200 µl PCR tube for each sample:

Reagent	**Volume (50 µl)**
DNA ligation product Step 3.14	8 µl
5X Phusion buffer	10 µl
dNTP mix (10 mM each dNTP)	1.5 µl
PCR primer (10 µM)	1 µl
Barcode primer (10 µM)	1 µl
water	28.5 µl

3.16 Centrifuge briefly and incubate in a thermal cycler using the following program:

Steps	**Temp / °C**	**Time**
Initial Denaturation	98	30 s
14 to 18 cycles (depends on the library concentration)	98	10 s
	65	30 s
	72	30 s
Final extension	72	5 min

3.17 Purify DNA using Zymo DNA Cleanup Kit. Elute in 20 µl water.

Size selection

3.18 Load the sample with loading buffer on 10% TBE Gel.

3.19 Run the gel at 100V constant voltage until the pink dye migrates to the bottom of the gel.

3.20 Dilute the SYBR gold in 1:10000 in TBE Buffer. Soak the gel in the staining buffer and incubate for 10 min at RT.

3.21 Cut out the smear as the size range of 150-400 bp on the gel.

3.22 Add 360 μl Gel Elution Buffer to the gel slice to elute the DNA. Incubate for 2 hours with rotation at RT.

3.23 Collect the supernatant by filtering out gel debris using a Costar column.

3.24 Add 1000 μl ethanol, 36 μl 3M NaAc and 1 μl GlycoBlue to the eluted library. Precipitate the DNA overnight at -20 °C or -80 °C for 15 min.

3.25 Centrifuge at 16000g for 15 min at 4 °C to pellet the DNA.

3.26 Wash the pellet with 1 ml ice-cold 70% ethanol twice.

3.27 Dissolve the DNA pellet in 10 μl water.

3.28 Measure DNA concentration using the Qubit dsDNA HS assay kit.

3.29 The library is ready for sequencing (See Note 2).

2.3.3 Notes

1. Multiple methods for shearing chromatin are available. Except for the probe sonicator described in the protocol, Bioruptor and the Covaris instruments can also be used. The detailed sonication settings should be optimized for different cells, according to the desired

final chromatin size and temperature controls. During sonication, high temperature should be avoided by adjusting the interval waiting time between each pulse.

2. All primers are designed for Illumina HiSeq2500 system. Sequencing as single-end 50 bp reads is usually sufficient and cost effective. The optimal sequencing depth depends on the number of R-loops of different targets. For HEK293T and K562 cells, 40 million reads per sample (at least 15 million unique mapped reads) is sufficient to generate a high confidence R-loop profile. Controls should be sequenced deeper than the R-ChIP samples (~50 million reads).

3. This protocol includes using the sonicated input as control to correct for the background of enriched R-loops. Besides this type of control, using irrelevant antibody "Normal IgG" can also provide good control for any bias introduced during the R-ChIP process.

2.4 R-loop features

We first used our R-ChIP data to determine the R-loop size in comparison with that directly visualized under EM. We found a median peak size of 199 bp by the narrow peak calling strategy of MACS2 [Feng et al., 2012] and a larger size (318 bp) by the broad peak calling strategy (Figure 2.8A). The vast majority (89.5%) of broad peaks encompassed narrow peaks (Figure 2.9A), and those uniquely identified by only one of the peak calling strategies were in general associated with weaker R-ChIP signals (Figure 2.9B). Therefore, the R-loop size range deduced by R-ChIP is similar to that observed under EM and by bisulfate sequencing [Duquette et al., 2004, Yu et al., 2003]. We next analyzed the sequence preference associated with R-loops. We

found strong G/C skew in the non-template DNA within mapped R-loop regions (Figure 2.8B). We further noted a peak in G distribution near the front of R-loops, with many containing 3–5 consecutive G residues (Figure 2.8B and 2.8C), consistent with the formation of G-quadruplex or some distorted structure to separate non-template from template DNA to promote R-loop formation [Duquette et al., 2004, Stork et al., 2016]. By comparing with the published G-quadruplex profile from in vitro folded DNA fragments [Chambers et al., 2015], we found a substantial overlap with our in vivo mapped R-loops, and the sequences that have the potential to form G-quadruplex were preferentially distributed in the non-template DNA strand (Figure 2.8D), which were also located at the front of mapped R-loops (Figure 2.8E). These data reveal a tight association of G-rich clusters in the non-template DNA with the potential of R-loop formation genome-wide.

In light of the remarkable agreement between in vivo R-loops mapped with R-ChIP and all biochemical properties known to promote R-loop formation and stabilization established in vitro, we next determined the genomic distribution of newly mapped R-loops. Although the association of R-loops with gene promoters has been a general consensus in the field, the existing R-loop maps also suggest their prevalent distribution in gene bodies and gene ends [Ginno et al., 2012, Ginno et al., 2013, Sanz et al., 2016, Stork et al., 2016]. We found that more than half of R-ChIP mapped peaks (7,649, 59.3%) resided in promoter proximal regions (±1 Kb from TSS) (Figure 2.8F). Additional R-loops were mapped to various locations within gene bodies (2,226, 17.2%), near gene terminal (846, 6.6%), or in intergenic regions (2,185, 16.9%). However, the average signal intensity at promoters was significantly higher than that in other genomic regions (Figure 2.8G). Therefore, despite strong potential of G-quadruplex formation in numerous regions in gene bodies, these observations suggest that the bulk of gene body sequences are

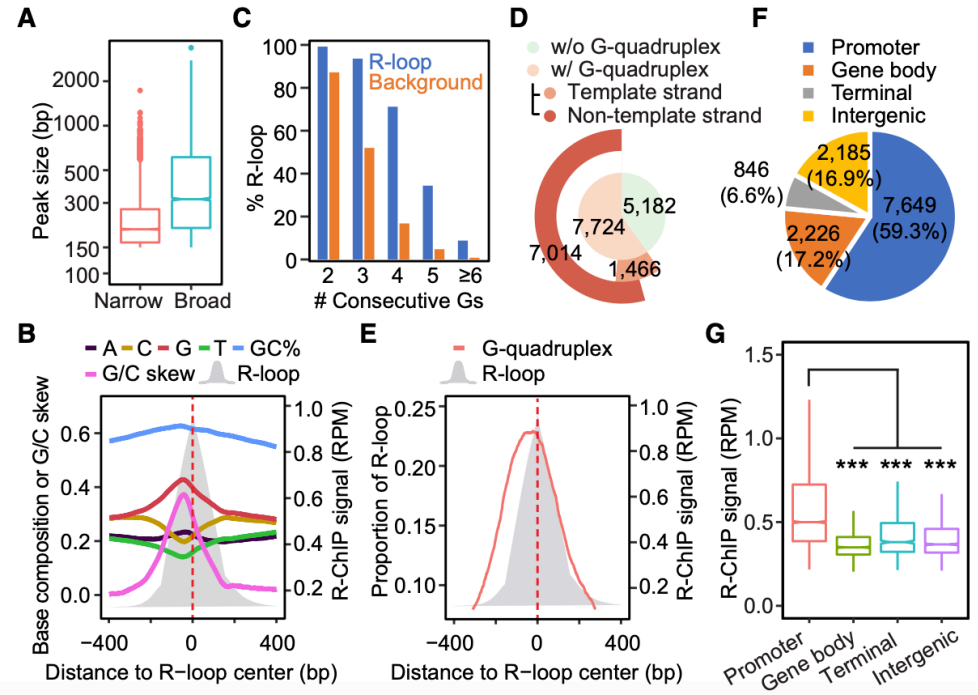

Figure 2.8: Sequence Features and Genomic Distribution of R-ChIP Signals. (A) The size distribution of R-ChIP peaks determined by the narrow or broad peak calling strategies of MACS2. (B) Base composition, G/C content, and G/C skew associated with a composite R-loop map. (C) Percentages of total R-loops according to associated consecutive G numbers (G-clusters) in the ±50 bp flanking region of the G/C skew summit in comparison with background. (D) Coincidence between R-ChIP mapped R-loops and potential G-quadruplex forming regions, emphasizing predominant overlap with G-quadruplex forming regions on the non-template DNA strand. (E) R-loop profile relative to sequences that have the potential to form G-quadruplex. (F) The genomic distribution of R-ChIP mapped R-loops. Various genomic regions are color coded according to the labels on the top. (G) The signal intensity distribution of R-ChIP peaks in different genomic regions.

prohibitive to R-loop formation, implying other critical requirements for efficient R-loop formation near TSSs. Overall, most detected R-loop peaks were coincident with open chromatin based on DNase I hypersensitivity, RNAPII occupancy, and multiple active chromatin marks (Figure 2.10), suggesting that active gene promoters are major hot spots for R-loop formation in the genome.

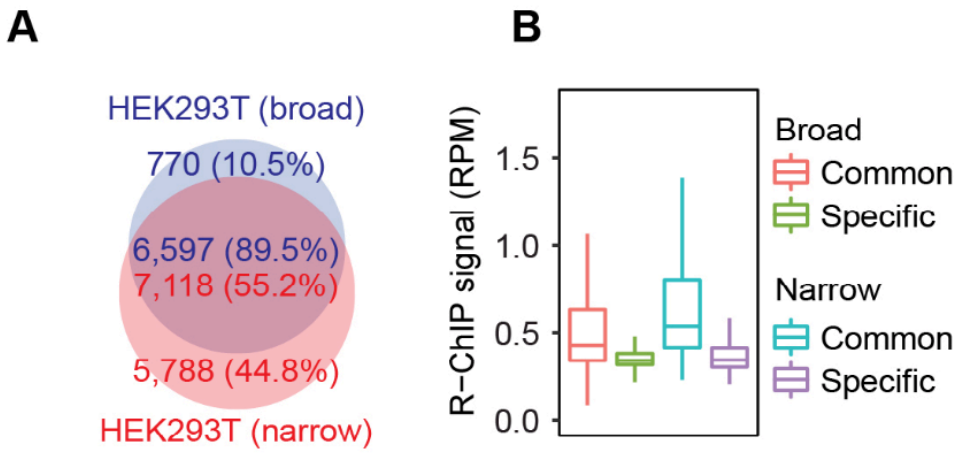

Figure 2.9: Comparison of R-ChIP Peaks Identified by Narrow and Broad Peak Calling Strategies.(A) Overlap of R-ChIP peaks in HEK293T cells expressing D210N mutant identified by narrow and broad peak calling strategies of MACS2. The number and the percentage of both overlapped and non-overlapped peaks are indicated. (B) Overall R-loop signal intensity associated with each of the four R-loop groups.

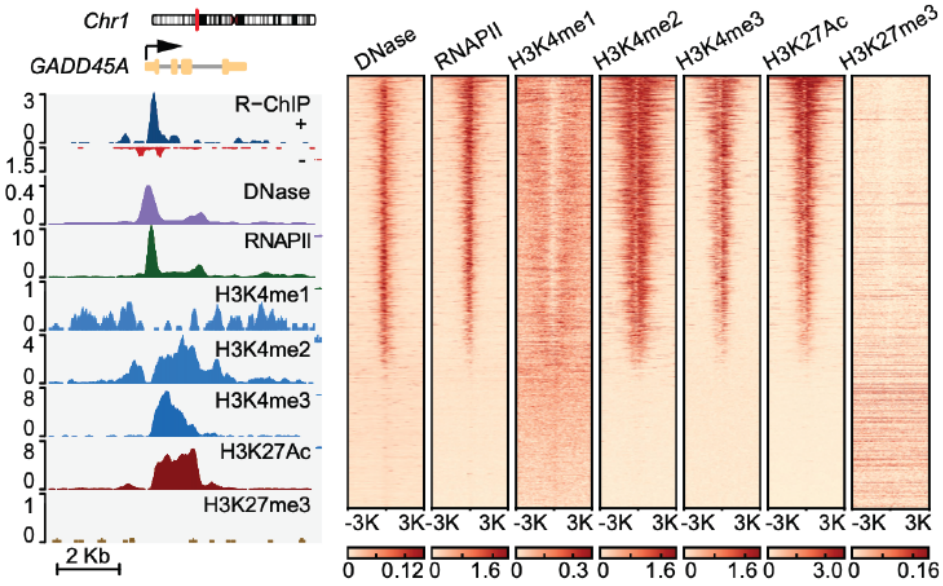

Figure 2.10: A representative genomic region covering the GADD45A gene locus, showing R-ChIP signals relative to open chromatin (DNase-seq), RNAPII occupancy, and various chromatin marks. The heatmap presentation of DNase-seq signals and ChIP-seq signals for RNAPII, H3K4me1, H3K4me2, H3K4me3, H3K27ac, and H3K27me3 in regions ±3 Kb from R-loop centers.

2.5 R-loop formation dynamics

2.5.1 R-loop induction in response to transcriptional perturbation

We next focused on understanding a puzzle about diminished, rather than increased, R-loop upon treating cells with the transcription inhibitor 5,6-dichloro-1-b-D-ribofuranosyl benzimidazole (DRB) [Sanz et al., 2016], a drug known to block transcriptional elongation in gene body but increase RNAPII pausing at TSS by inhibiting the RNAPII C-terminal domain (CTD) kinase pTEFb [Jonkers and Lis, 2015]. We first performed R-ChIP-qPCR on several gene promoters upon DRB treatment (2 hr), observing elevated R-loops in all cases (Figure 2.11A). Careful comparison between our data and those reported earlier [Sanz et al., 2016] revealed that the previous study placed PCR primers on DRIPc-seq mapped R-loop peaks, which were 1.5 Kb away from each promoter, whereas our PCR primers were designed to interrogate individual gene promoters according to R-ChIP mapped R-loop peaks at TSSs. This explains diminished R-loop signals from the previous study because DRB is known to block RNAPII elongation beyond major pausing sites near TSSs.

This observation prompted us to further investigate R-loop dynamics in relationship to transcriptional pause release after washing away DRB. On 11 gene promoters, we performed ChIP-qPCR for RNAPII and R-ChIP-qPCR to detect R-loop decay at different time points up to 72 min after removing DRB (Figure 2.11B). Interestingly, we found that RNAPII binding decreased ahead of R-loop resolution in all cases, as indicated by differential decay rate on individual gene promoters as well as the averaged curves (thick blue lines) based on all 11 genes we surveyed (Figure 2.11B). Encouraged by these findings, we further extended the analysis globally

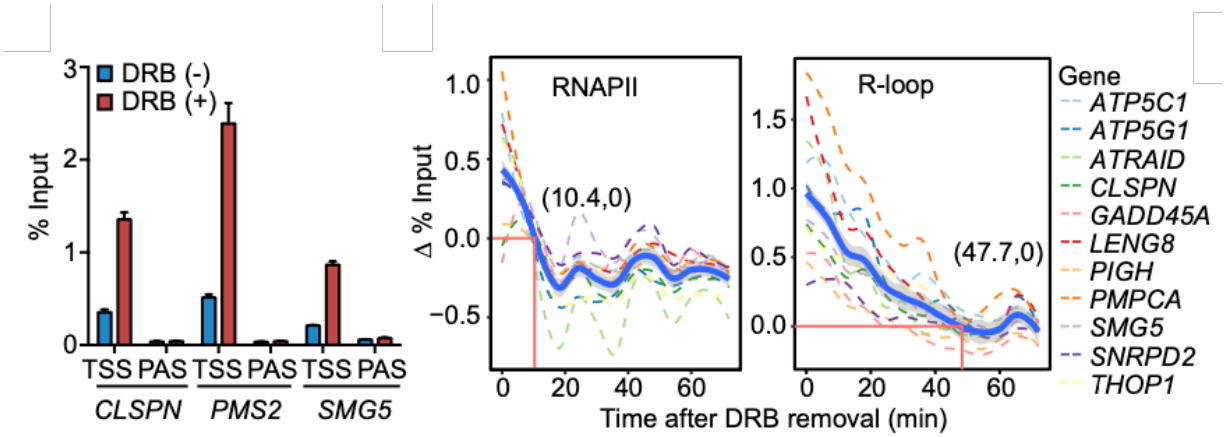

Figure 2.11: Induction of TSS-associated R-loops upon DRB treatment (2hr) on three representative genes detected by R-ChIP-qPCR. Results were calculated as the percentage of input and presented as mean ± SEM (n = 3 technical replicates). Dynamics of the RNAPII occupancy and R-loop level following DRB removal at TSS regions of 11 representative genes by RNAPII ChIP-qPCR and R-ChIP-qPCR. Cells were first treated with DRB for 2 hr and then collected every 6 min after DRB removal. Thick blue line: average RNAPII ChIP-qPCR and R-ChIP-qPCR values of 11 genes. Red lines indicate the time point, at which the average RNAPII occupancy or R-loop level returned to the baseline level in untreated cells. The associated numbers indicate the time (min) for returning to the baseline.

by generating highly reproducible R-ChIP libraries on mock-treated HEK293T cells [DRB(–)] or cells treated with DRB for 2 hr [DRB(+)] or after DRB removal for 30 min [post-DRB] (Figure 2.14A). As shown on a specific example (SAE1) (Figre 2.5), DRB treatment induced R-loop at TSS and DRB removal returned the R-loop to the original level (Figure 2.12C). We also performed GRO-seq to monitor transcriptionally engaged RNAPII under these conditions, confirming DRB-induced pausing at TSS and resumed RNAPII elongation after DRB removal, as indicated by reduced nascent RNA signals at TSS and restored signals in gene body (Figure 2.12D). These trends held for all expressed genes from global analysis (Figure 2.13E and 2.13F). As demonstrated on the 11 genes surveyed by R-ChIP-qPCR (Figure 2.14B), we also observed a large amount of remaining R-ChIP signals compared to largely cleared GRO-seq signals from TSSs (Figure 2.14C). These data suggest that, while R-loop induction is coupled with RNAPII

pausing at TSSs, R-loop resolution is not required for RNAPII pause release (note that most remaining R-loops were unlikely to result from newly recruited RNAPII because of small amounts of nascent RNAs generated after DRB removal, see Figure 2.12D).

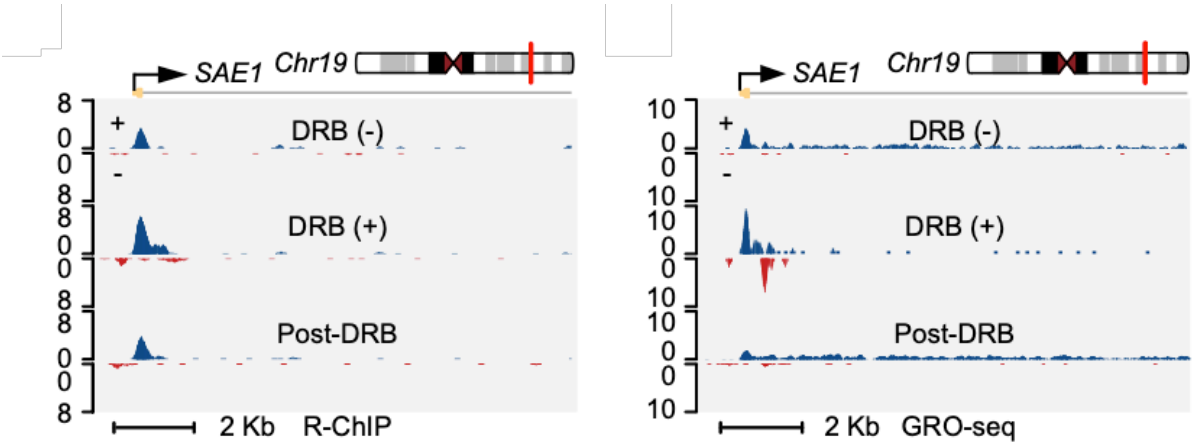

Figure 2.12: A representative genomic region covering the TSS region of SAE1, showing R-ChIP (C) and GRO-seq (D) signals in response to DRB treatment [DRB(+)] and removal (Post-DRB).

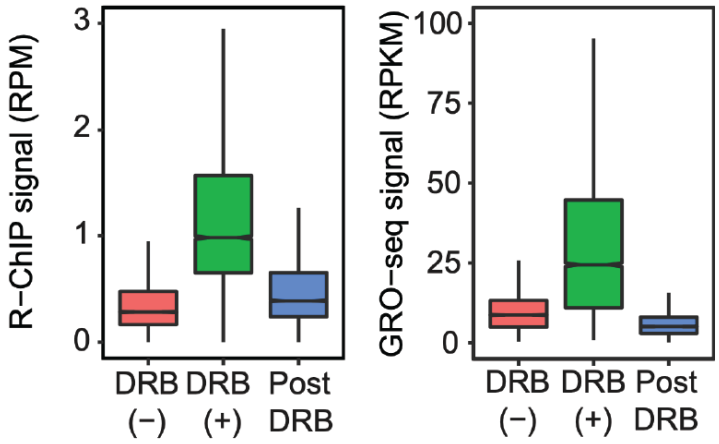

Figure 2.13: Signal intensity distribution of overall R-loop levels detected by R-ChIP (E) and RNAPII activities by GRO-seq (F) at TSSs in response to DRB treatment [DRB(+)] and removal (Post-DRB).

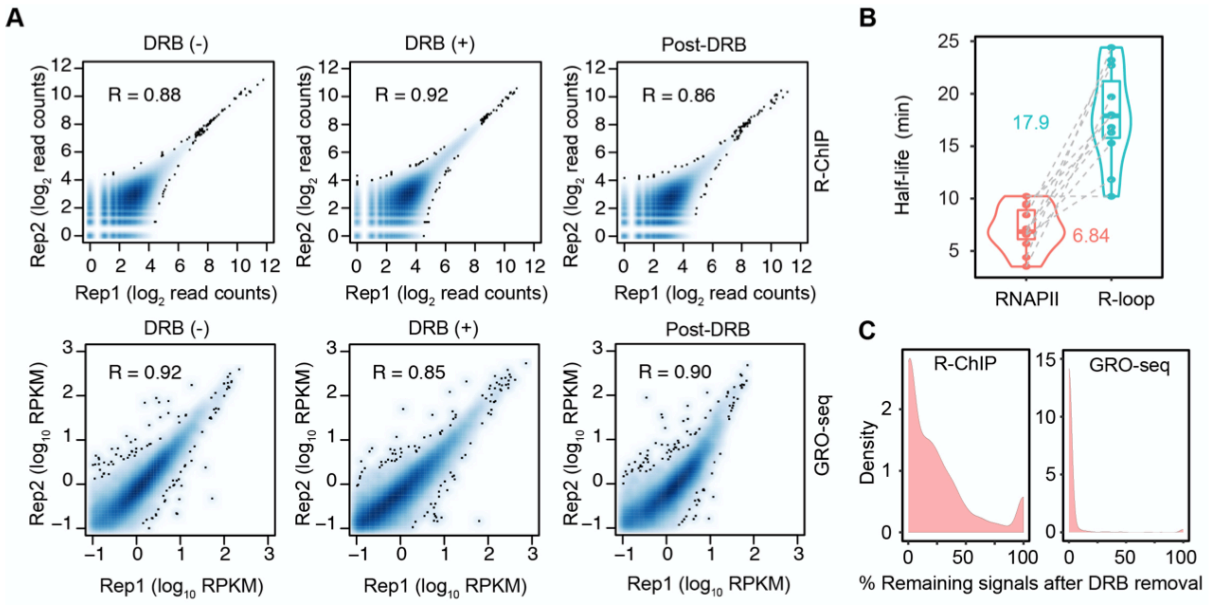

Figure 2.14: R-loop Dynamics in Response to DRB Treatment. (A) Pairwise comparison of two independent R-ChIP (top panels) and GRO-seq (bottom panels) replicates under three different DRB treatment conditions as indicated. Reads coverage of each continuous 3 Kb bin across the human genome was calculated for each R-ChIP dataset. The gene expression level in terms of RPKM for each transcript was calculated for each GRO-seq dataset. Pearson correlation coefficients were computed to evaluate the reproducibility. (B) The half-life for RNAPII and R-loop decline at the TSS region of 11 examined genes following DRB removal. Numbers indicate the median half-life. (C) The distribution of remaining R-ChIP detected R-loop and GRO-seq signals after DRB removal (30min).

2.5.2 A free RNA end required to promote R-loop formation

During our analysis of R-ChIP signals, we curiously noted that the majority of R-ChIP mapped R-loops were each associated with a nearby free RNA end (Figure 2.15). This makes sense from topological consideration because a free RNA end would enable efficient RNA invasion into duplexed DNA behind an elongating RNA polymerase. To directly determine the requirement for such free RNA end to promote R-loop formation, we engineered a ribozyme to co-transcriptionally generate a free RNA end within a transcription unit driven by the CMV promoter and terminated by a build-in poly(A) site in the vector (Figure 2.15). As a control, we

introduced a point mutation in the ribozyme, which compromises both the activity and specificity of the ribozyme as characterized previously [Fong et al., 2009]. Because R-loop formation also requires a critical sequence context, we selected an R-loop-promoting G-rich sequence characterized earlier [Zhang et al., 2014] and another R-loop-promoting sequence from one of the endogenous genes (CPSF7) we identified from our R-ChIP mapping data and separately inserted them into the plasmid downstream of the ribozyme to mimic the R-loop prone sequence context.

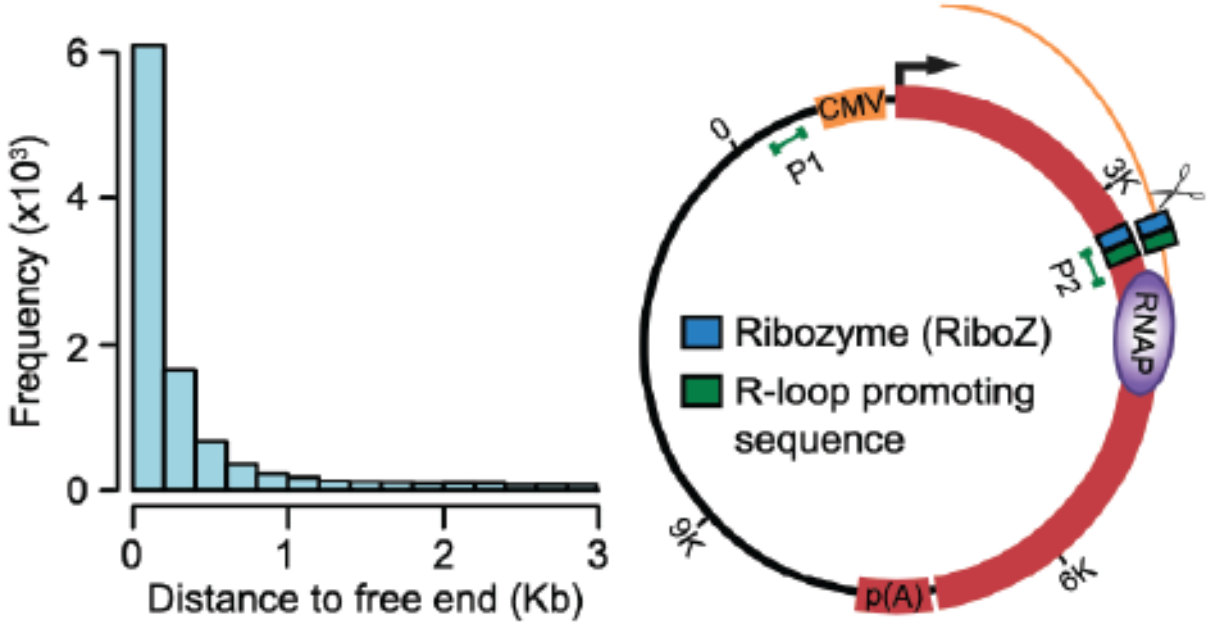

Figure 2.15: The distance distribution of known free RNA ends relative to R-ChIP mapped R-loops. The R-loop reporter plasmid: WT or mutant hepatitis d ribozyme with or without a R-loop promoting sequence were cloned into 2.6 Kb downstream of the CMV promoter in a pcDNA5-based expression vector carrying part of the luciferase gene fused to the 3-UTR of the FUBP1 gene. Two pairs of primers targeting a promoter upstream region (P1) and the potential R-loop forming region (P2) were used for R-ChIP and DRIP analyses, as indicated.

Based on this experimental design, we generated six constructs containing the WT or mutant ribozyme with or without R-loop-promoting sequences. Given that plasmids bearing pro-R-loop features could efficiently produce R-loops in transfected cells, which were directly

visible under EM [Duquette et al., 2004], we transfected these constructs into HEK293T cells expressing the catalytically dead RNASEH1 to capture R-loops by R-ChIP followed by real-time PCR analysis of a promoter upstream region (P1, as a negative control) and the predicted R-loop forming region (P2). We found that, without inserting any R-loop-promoting sequence, neither WT nor mutant ribozyme was sufficient to trigger R-loop formation (Figure 2.16, upper panel). After inserting the G-rich sequence (Figure 2.16, middle panel) or the R-loop-promoting sequence from CPSF7 (Figure 2.16, bottom panel), we detected the induced R-loop on both constructs containing the WT ribozyme. The level of R-loops was significantly reduced, but not diminished, with the mutant ribozyme (Figure 2.16), likely due to aberrant RNA fragmentation activities of the mutant ribozyme as reported earlier [Fong et al., 2009].

To validate the detected R-loops, we wished to use an entirely independent strategy for R-loop detection. Although DRIP suffers from poor resolution in genome-wide analysis of R-loops, it is suitable for detecting R-loops in specific loci. We therefore transfected the plasmids into HEK293T cells and subjected purified DNA to RNase H treatment followed by IP with the S9.6 antibody. To control for non-specific binding on S9.6 beads, we performed a parallel spike-in experiment with purified DNA from mock-transfected cells mixed with individual plasmids. By first normalizing the total input followed by subtracting the background IP signals (Figure 2.19), we demonstrated RNase H-sensitive R-loop that was only detectable at the ribozyme inserted region on both plasmid reporters (Figure 2.17). These data not only demonstrate the requirement of a free 50 RNA end for inducing R-loop formation, but also suggest a revised model for R-loop initiation and elongation, as illustrated in Figure 2.18.

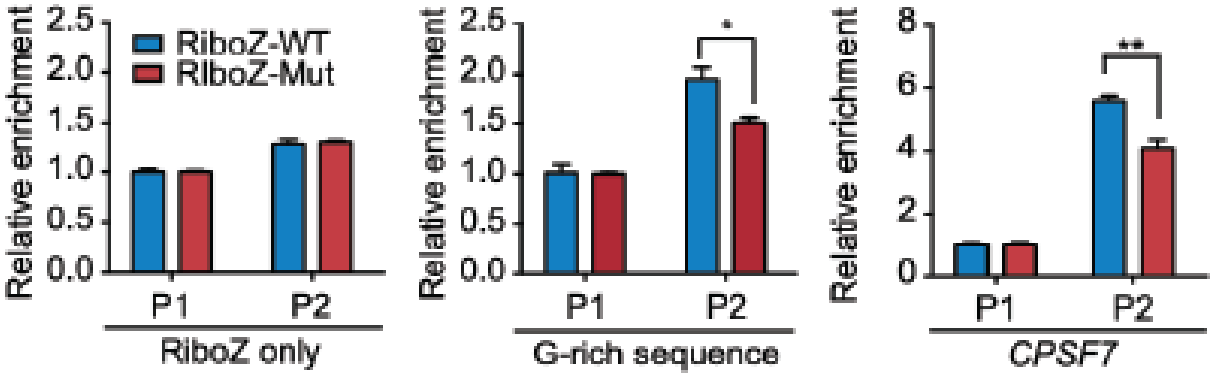

Figure 2.16: R-ChIP-qPCR results on RNASEH1/D210N expressing HEK293T cells transfected with plasmid containing WT or mutant ribozyme without any R-loop-promoting sequence (top), with a G-rich R-loop-promoting sequence (middle), or with a R-loop-promoting sequence from CPSF7 (bottom). Results were calculated as fold enrichment of signals at the P2 region relative to total input and then normalized against control signals from the P1 region. Data are presented as mean ± SEM (n = 4 technical replicates). *p < 0.05; **p < 0.01, unpaired Student's t test.

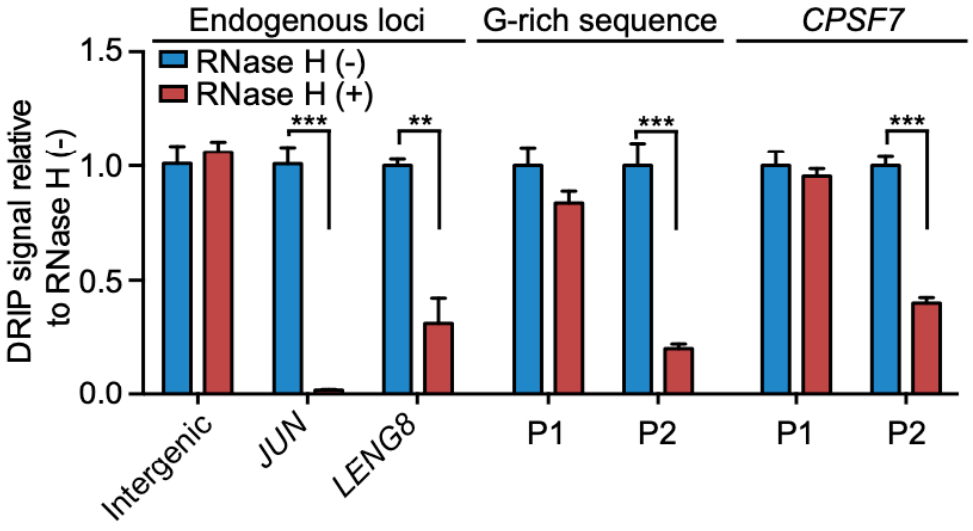

Figure 2.17: DRIP-qPCR analysis on HEK293T cells transfected with plasmid containing WT ribozyme and individual R-loop-promoting sequences. Purified DNA from each sample was mock-treated or treated with RNase H before DRIP. Examined are one intergenic control region and two endogenous. R-loop prone gene promoters identified by R-ChIP, as well as P1 and P2 regions on individual transfected plasmids. Results were calculated as relative DRIP-qPCR signals after setting signals from mock-treated samples as 1 and presented as mean ± SEM (n = 4 or 5 technical replicates). **p < 0.01; ***p < 0.001, unpaired Student's t test.

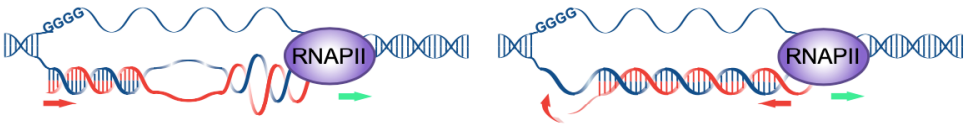

Figure 2.18: Current and revised models for R-loop formation and elongation.

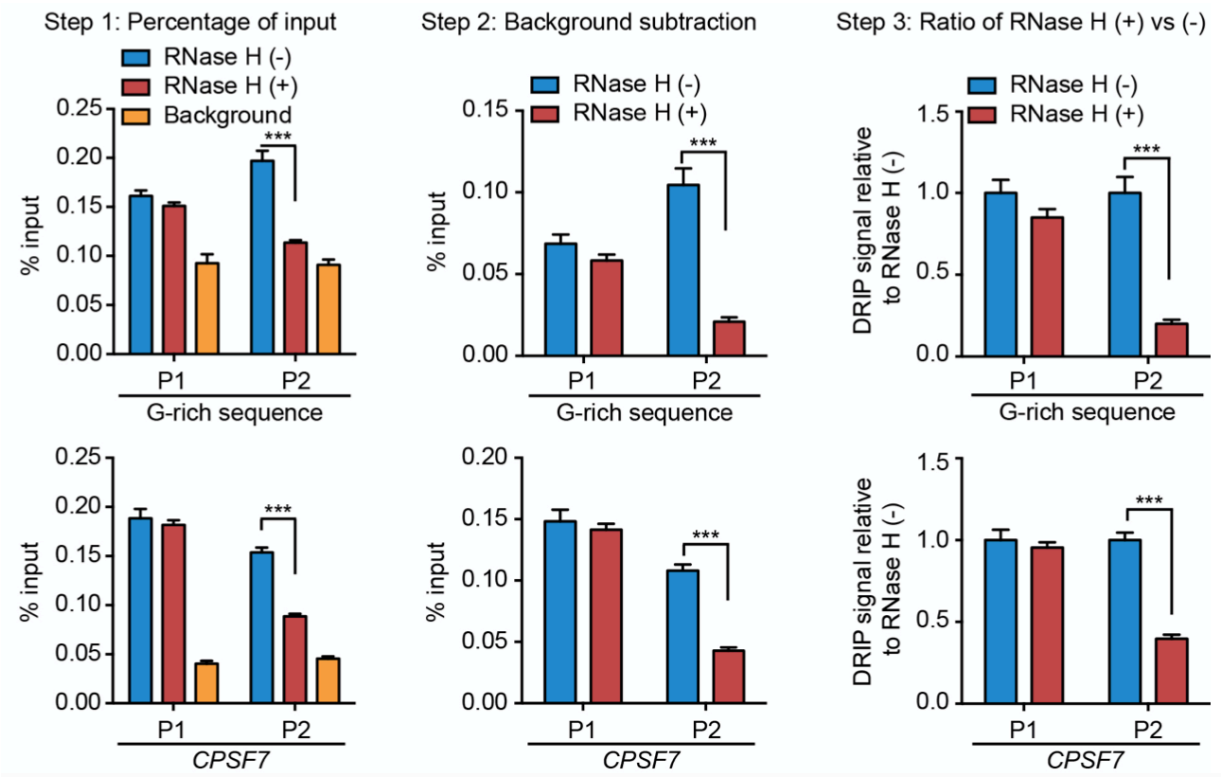

Figure 2.19: Step-wise processing and normalization of DRIP-qPCR results. (A) DRIP-qPCR results were first calculated as the percentage of total input. Background was determined by using the plasmid spike-in sample. Data were presented as mean ± SEM (n = 5 technical replicates). ***p < 0.001 based on unpaired Student's t-test. (B) DRIP-qPCR results after background subtraction. (C) Relative enrichment obtained by calculating the ratio of DRIP signals mock-treated or treated with RNase H.

2.6 Comparison with other R-loop mapping methods

2.6.1 Comparison with S9.6-based methods

Early efforts used indirect footprint approaches to locate R-loop-forming regions. In principle, singlestranded DNA (ssDNA) in the R-loop structure can be modified at cytosine residues by bisulfite under non-denaturing conditions, and several early studies thus combined bisulfite conversion and Sanger sequencing to map R-loops at specific genomic loci of interest [Li and Manley, 2005, García-Benítez et al., 2017].

DNA/RNA immunoprecipitation sequencing (DRIP-seq) is the first and most widely used technique developed for genome-wide capture of R-loops by using fragmented chromatin for IP with S9.6 antibody, followed by sequencing of the recovered R-loop-containing DNA fragments8. Similar approaches have been applied to multiple biological systems, including yeast [Wahba et al., 2016], plants [Xu et al., 2017b] and mammals [Chen et al., 2015b]. To further improve the resolution and robustness of DRIP-seq, a number of strategies have been developed to sequence RNA, or template DNA in S9.6-captured material [Chen et al., 2015b]. Other efforts to increase the resolution, specificity or sensitivity of DRIP-seq-based approaches have also been made through combining DRIP with bisulfite footprinting to identify R-loop-associated ssDNA (bisDRIP-seq), or via S1 nuclease digestion to remove non-template ssDNA in R-loop regions before sonication in order to prevent its re-annealing back to template DNA during IP (S1-DRIP-seq) [Wahba et al., 2016]. It remains unclear to what extent sonication may disrupt fragile R-loops, especially with unfixed cells.

DRIP-seq and its derivatives rely on the specificity of the S9.6 antibody, which has been

recently questioned [Vanoosthuyse, 2018, Hartono et al., 2018]. S9.6 predominantly recognizes RNA–DNA hybrids in a sequence-independent manner. However, it also binds dsRNA with a lower affinity than RNA–DNA hybrids, leading to substantial false-positive signals. This is because nascent RNA transcripts, part of which may be folded into dsRNA, are known to extensively associate with transcribing DNA in the nucleus [Li et al., 2017, Bell et al., 2018], and thus, such chromatin-tethered RNAs would give rise to false-positive signals. A key control in both DRIP-seq and DRIPc-seq is the treatment of the same sample before IP with purified bacterial RNase H, to digest RNA within R-loops. However, caution must be taken in interpreting data from such 'control'. Owing to biased restriction digestion of DNA or insufficient RNA fragmentation, DRIP-seq and DRIPc-seq may capture not only R-loops, but also their associated DNA and RNA fragments [Halász et al., 2017]. In DRIPc-seq, for example, a portion of an RNA may be engaged in R-loop formation, whereas other parts of the RNA remain as unengaged RNA that contains both singleand double-stranded regions. Once the anchor in the R-loop is removed by RNase H treatment, all associated RNA would be lost. Therefore, it would be difficult to conclude that all RNase H-sensitive signals correspond exactly to the R-loop formation regions. The S9.6 antibody may also capture any dsRNA or dsRNA-containing RNA anchored to DNA via triplex formation [Schmitz et al., 2010, Bacolla et al., 2015, Mondal et al., 2015, O'Leary et al., 2015, Postepska-Igielska et al., 2015, Li et al., 2016] or other unknown mechanisms. This is supported by a recent report showing that DRIPc-seq detects a substantial fraction of RNase H-resistant signals that are instead sensitive to a dsRNA-specific endoribonuclease (RNase III) in yeast [Hartono et al., 2018].

A thorough comparison of R-ChIP and S9.6-based methods, including DRIP-seq, DRIPc-seq and RDIP-seq, has been performed as reported [Chen et al., 2017]. R-ChIP-mapped R-loops

showed high resolution and accordance with all known sequence features of R-loop-forming regions, suggesting improved accuracy and specificity. Although both R-ChIP and S9.6-based methods revealed enrichment of R-loops at transcription start sites (TSSs), the major discrepancy is highly enriched signals in both gene bodies and transcription termination sites (TTSs) detected by S9.6-based methods, but not by R-ChIP. The theoretical basis for such discrepancy is currently unknown. A formal possibility is that R-ChIP captures the action sites of RNASEH1 in the genome, but not in other regions, such as gene bodies and gene ends.

Because of the major discrepancy between the existing R-loop mapping methods, it is important to develop technologies independent of S9.6 or RNASEH1. The problem is that, currently, there is no 'gold standard' for detecting R-loops. The recently developed bisDRIP-seq technology combines S9.6 capture with bisulfite sequencing [Dumelie and Jaffrey, 2017], which is appealing on the basis of the principle of its experimental design. Probably because of the noisy nature of the data, the authors had mainly used ensemble data for analysis after combining multiple independently generated libraries. We used the same ensemble bisDRIP-seq data to compare with the signals captured by R-ChIP versus those detected by DRIP-seq and DRIPc-seq. On the basis of the meta-gene analysis, both R-ChIP and bisDRIP-seq detected R-loops at TSSs, but few in gene bodies or ends, which is in contrast to strong signals detected by both DRIP-seq and DRIPc-seq (Figure 2.20). A high consistency between R-ChIP and bisDRIP-seq is further illustrated by compiling bisDRIP-seq signals on mapped genomic regions with different methods, showing that the bisDRIP-seq signals are highly enriched at the peak regions detected by R-ChIP, but not by DRIPc-seq and DRIP-seq (Figure 2.20). On specific gene examples, such as NEAT1, a broadly expressed long noncoding RNA, R-loop signals were detected by all four methods, but the signals detected by both R-ChIP and bisDRIP-seq were narrowly enriched near its TSS,

whereas much broader signals were detected by both DRIP-seq and DRIPc-seq (Figure 2.20, top left). For DNTTIP1, signals in a broad region of the gene body were detected by DRIPc-seq and less so by DRIP-seq, but no specific signal was detected by either R-ChIP or bisDRIP-seq (Figure 2.20, top right). By contrast, for RPPH1 and HIST1H2BG, the two gene loci previously documented by bisDRIP-seq, R-ChIP, but not DRIPc-seq or DRIP-seq, showed the same signals as bisDRIP-seq (Figure 2.20, bottom panels). Together, these comparisons suggest a high degree of agreement between R-ChIP and bisDRIP-seq.

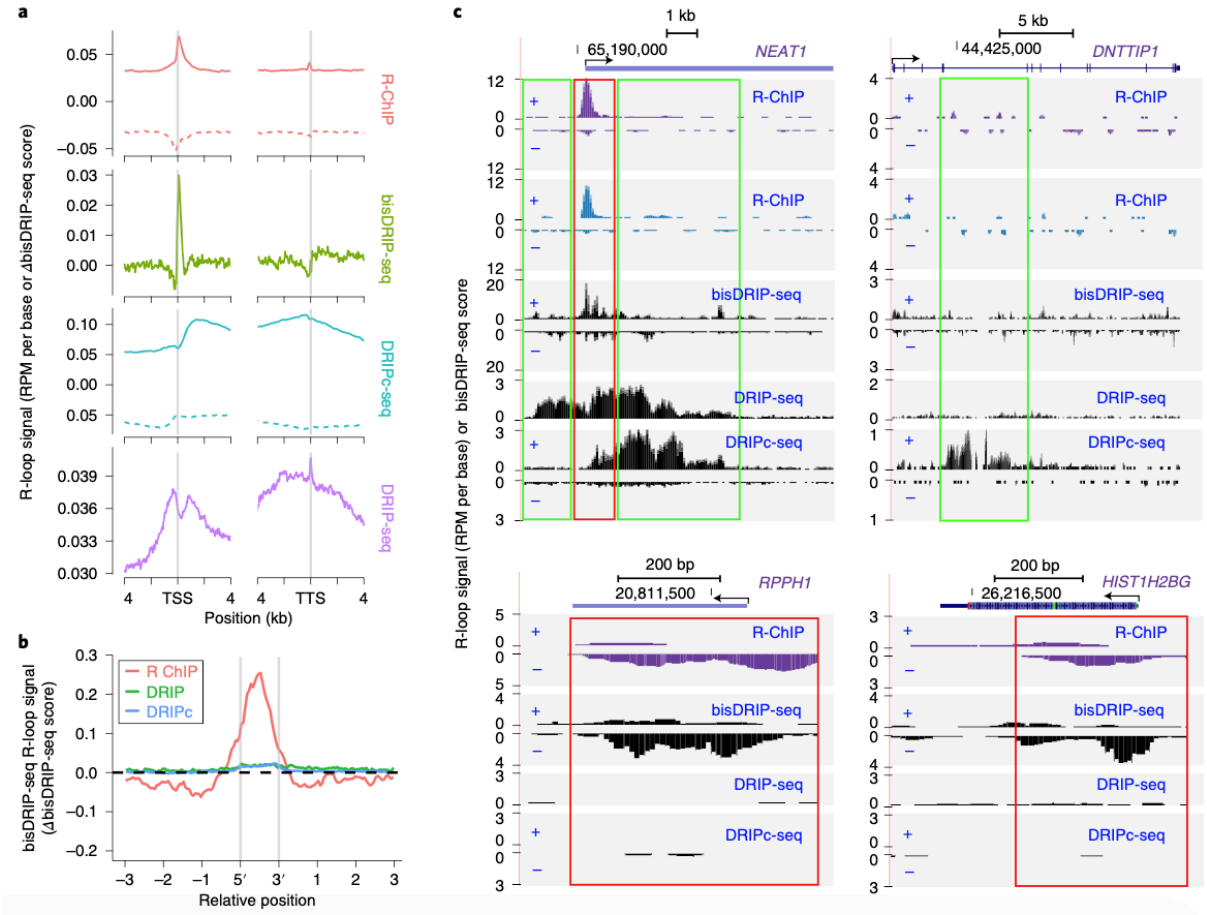

Figure 2.20: Comparison of R-ChIP and S9.6-based methods. a, Meta-gene analysis of signals detected by R-ChIP, bisDRIP-seq, DRIPc-seq and DRIP-seq at the ±4-kb region of TSS and TTS, respectively. Note that y-axis units are different for individual methods. R-loop signals are measured in terms of reads per million (RPM) per base by R-ChIP, DRIPc-seq and DRIP-seq, or the difference between template and non-template strand bisDRIP-seq scores by bisDRIP-seq . bisDRIP-seq signals above or below 0 indicate the existence of R-loops on the forward or reverse strand. For R-ChIP and DRIPc-seq, solid and dotted lines represent potential R-loop signals from the forward and reverse strands, respectively. b, bisDRIP-seq R-loop signals at peak regions defined by R-ChIP, DRIPc-seq (DRIPc) and DRIP-seq (DRIP) data. c, Signals detected by R-ChIP (purple for K562 cells and blue for HEK293T cells), bisDRIP-seq, DRIPc-seq and DRIP-seq at representative genomic loci. Regions with consistent signal profiles between R-ChIP and bisDRIP-seq are highlighted by red rectangles, whereas those detected only by DRIP-seq or DRIPc-seq are highlighted by green rectangles. TSS, transcription start site; TTS, transcription termination site.

2.6.2 Comparison with RNaseH1-based methods

In general, R-loops could be recognized by using either a specific monoclonal antibody (S9.6) or catalytically inactive RNaseH1 [Vanoosthuyse, 2018]. The most original S9.6-based method to capture R-loop containing DNA for deep sequencing is DRIP-seq (DNA-RNA immunoprecipitation sequencing) [Ginno et al., 2012]. This strategy has been further refined to improve the specificity and resolution, as represented by ssDRIP-seq, DRIPc-seq and bisDRIP-seq [Xu et al., 2017b, Sanz et al., 2016, Dumelie and Jaffrey, 2017]. A major concern on S9.6-based approaches is related to the binding specificity of the antibody, as it also has significant affinity for double-stranded RNAs, which may contribute to explanation of some major discrepancies in literature [König et al., 2017]. Alternatively, R-ChIP (R-loop mapping by chromatin immunoprecipitation) was developed by taking advantage of the high binding affinity of RNaseH1 for DNA/RNA hybrids [Chen et al., 2017]. In this technology, the expression of a catalytically inactivated version of this enzyme in cells enables efficient capture of R-loops for deep sequencing by using a standard ChIP protocol.

Like the DRIP-based method, the R-ChIP method has been evolved to multiple versions. For example, RR-ChIP sequences RNAs associated with RNaseH1-captured R-loops [Tan-Wong et al., 2019]. A major shortcoming for both R-ChIP and RR-ChIP is the requirement to express a mutant RNaseH1 in cells, which may not be feasible in certain applications and/or cause potential artifacts due to overexpressed RNaseH1 that may be dominant over the endogenous enzyme. These concerns have prompted the development of additional RNaseH1-based methods (Figure 2.21). MapR was developed by fusing mutant RNaseH1 with a general DNA cutting enzyme MNase [Yan and Sarma, 2020]. BisMapR is a derivative of MapR for constructing strand-specific

libraries [Wulfridge and Sarma, 2021]. These technologies essentially use RNaseH1 to guide MNase to cleave DNA in the genome where R-loops are located. R-loop CUT&Tag combines inactive RNaseH1 with the CUT&Tag sequencing technique to map R-loops by using the transposase Tn5 for simultaneous DNA cutting and primer ligation [Wang et al., 2021]. However, in general, these techniques generate much fewer peaks compare to R-ChIP, which may result from compromised affinity and specificity of purified RNaseH1 in the absence of potential co-factors, such as the single strand DNA binding protein RPA, involved in enhancing RNaseH1 targeting to R-loops in mammalian cells [Nguyen et al., 2017].

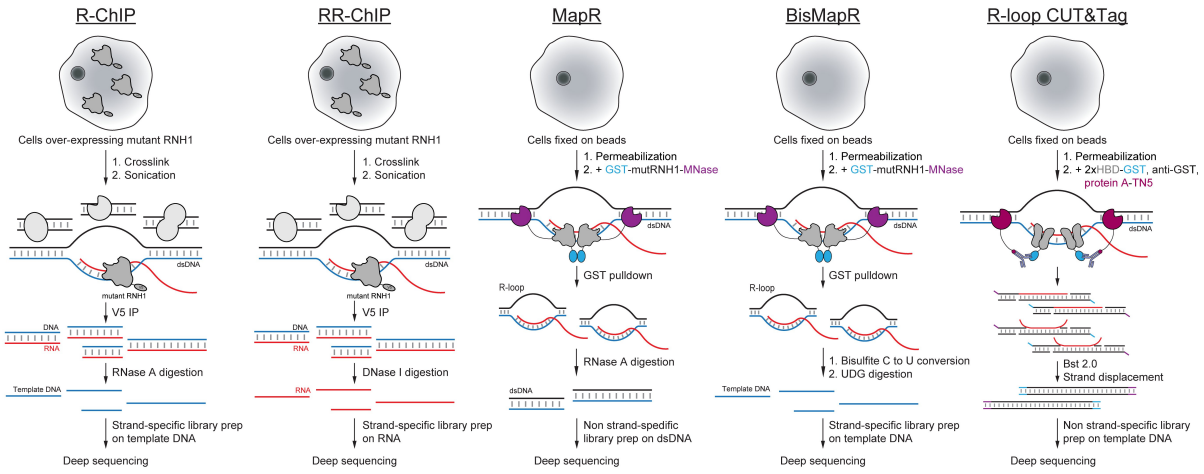

Figure 2.21: Schematic of R-loop methods based on RNaseH1. In R-ChIP and RR-ChIP, a cell line is first constructed to express a catalytically inactive RNaseH1, and R-loops are captured using a ChIP-seq protocol. In R-ChIP the template DNA is used for strand-specific library construction; in RR-ChIP, RNA enriched in R-loops is isolated for library construction. In MapR, BisMapR and R-loop CUT&Tag, cells are fixed on beads followed by permeabilization. In MapR, R-loops are cut and released by GST tagged inactivated RNaseH1 fused to MNase and released R-loops are enriched by GST pulldown. In BisMapR, captured R-loops are treated with bisulfite to convert C to U on the ssDNA. RNA within R-loops are removed by RNaseH1 and second strand synthesis is performed in the presence of dUTP. After UDG digestion, the remaining template DNA is used to construct a strand-specific library. In R-loop CUT&Tag, two hybrid binding domains of RNase H1 fused to GST (2xHBD-GST) are used to bind R-loops. Anti-GST linked to Tn5 is next used to cut and tag R-loops as in the CUT&Tag protocol.

Each R-loop mapping strategy may on one hand improve the existing technologies and

on the other hand introduce additional shortcomings, thus leading to significant discrepancies in literature. For example, R-ChIP predominately detects signals at gene promoters; however, RR-ChIP additionally generates abundant signals in gene bodies (Figure 2.22). Such extra gene-body signals may result from chromatin associated RNAs clustered around gene promoters. While MapR and R-loop CUT&Tag overcome the need to express an exogenous mutant RNaseH1, both show extensive signals in intergenic regions (Figure 2.22), which may result from nonspecific targeting of purified RNaseH1. Judging which signals represent bona fide R-loops is a quite challenging problem because of the lack of a "gold standard" for mapping genuine R-loops in the genome.

Because a key feature of R-loops is the single-stranded DNA replaced from the double helix, the ability to detect such single-stranded DNA may be leveraged to compare between different R-loop mapping methods. However, it is important to keep in mind that detectable genomic regions associated with single-stranded DNA do not necessarily represent R-loop formation activities. According to this rationale, we utilized the data generated by KAS-seq (Kethoxal-Assisted Single-stranded DNA sequencing), a recent developed technique for mapping single-stranded DNA along the genome [Wu et al., 2020], to compare between RNaseH1-based methods. It appears that the signals detected with KAS-seq show the highest coincident with those detected with R-ChIP Figure 2.22.

Given the robust correlation between the data generated by R-ChIP and KAS-seq, we believe that R-ChIP still represents one of the most robust R-loop mapping strategies developed to date.

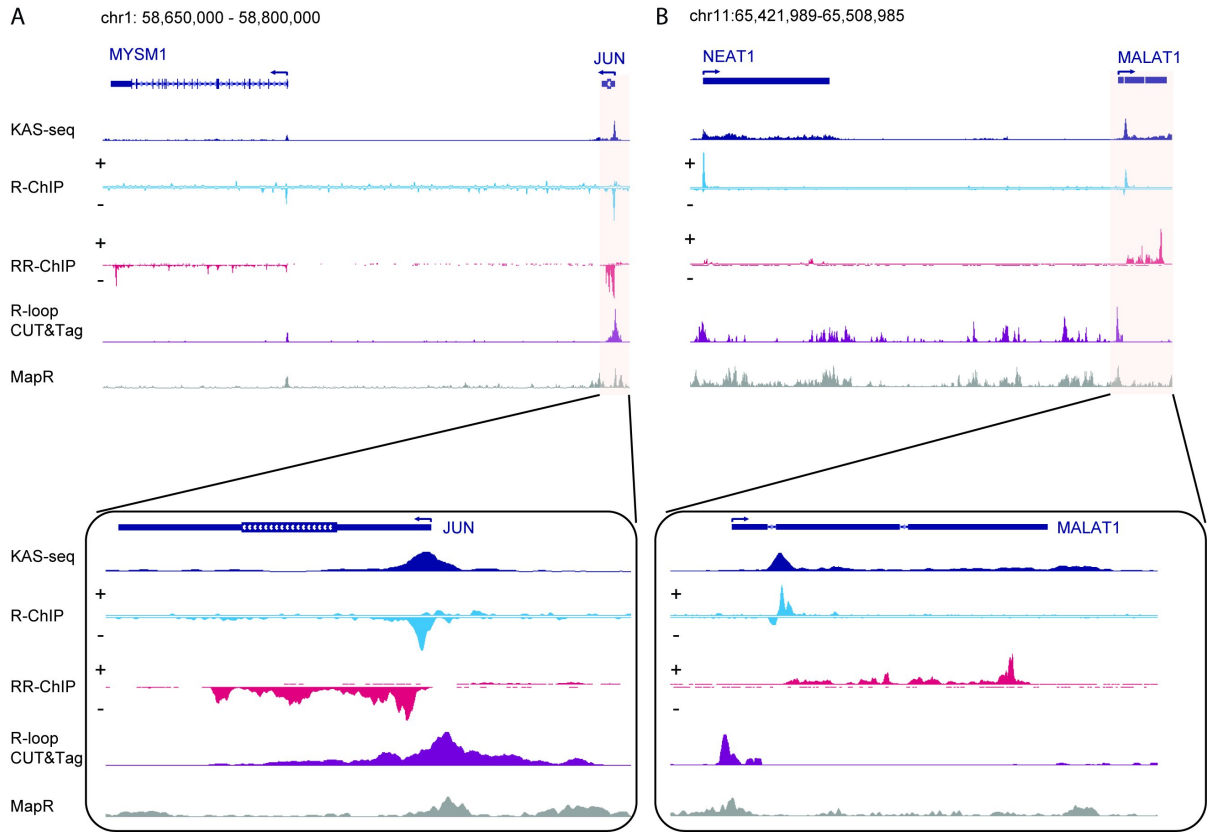

Figure 2.22: Genome browser view of R-loop signals detected by different methods. A. The genomic region containing JUN and MYSM1. B. The genomic region containing NEAT1 and MALAT1. Both panels display R-loop signals (normalized to reads per million, RPM) detected by R-ChIP, RR-ChIP, R-loop CUT&Tag and MapR. KAS-seq signals are provided for comparison on top of each panel.

2.7 Acknowledgements

Chapter 2, in part, are reprints of the material as it appears in X. Zhang, Y. Hao, X.-D. Fu,"Mapping R-loops using catalytically inactive RNaseH1 (R-ChIP)", *Methods in Molecular Biology*, 2021, J.-Y. Chen, X. Zhang, X.-D. Fu, and L. Chen., "R-Chip for genome-wide mapping of r-loops by using catalytically inactive RNASEH1.", *Nature Protocols,*, 2019, L. Chen, J.-Y. Chen, X. Zhang, Y. Gu, R. Xiao, C. Shao, P. Tang, H. Qian, D. Luo, H. Li, Y. Zhou, D.-E. Zhang, and X.-D. Fu. "R-chip using inactive RNase H reveals dynamic coupling of r-loops with

transcriptional pausing at gene promoters.“ *Molecular cell*, 2017, The dissertation author was the primary investigator and author of this paper.

Chapter 3

CSB Resolves R-loops by resuming stalled Pol II at poly T tract

3.1 Introduction

R-loops, by-products of transcription, are widely believed as a source of DNA damages due to the exposed single strand DNA (ssDNA). Double strand breaks (DSB) are induced when R-loops triggered transcription-coupled nucleotide excision repair (TC-NER), a major type of transcription-coupled repair (TCR) [Hanawalt and Spivak, 2008]. Cockayne syndrome group B (CSB) is proved to be a key regulator contributed to the genome instability in the development from excess R-loops to DNA damage sites [Sollier et al., 2014]. However, recent study shows that CSB is recruited by R-loops to repair the DSB site by initiating another type of TCR pathway, transcription-coupled homologous recombination (TC-HR) [Teng et al., 2018]. Therefore,

digging out how CSB interact with R-loops is crucial for understanding the "dual" role of CSB in R-loop processing, either to repair DSB or induce more DSB during transcription.

While R-loops are potent active regulators in TCR, they are considered as toxic by-products for decades to induce DNA damages, when the excess R-loops trigger TC-NER with the help of CSB[Sollier et al., 2014]. Once TC-NER initiated at R-loops, more DSB would be induced around R-loops due to the excision activity of XPG and XPF, two key factors in TC-NER. This has been long thought as the mechanism of how unscheduled R-loops develop into DNA damage. It remains unclear that whether R-loops are required to trigger the TCR, or R-loops is just a by-product during TCR initiation. Because CSB is the initiation factor of multiple TCR pathways, it is important to study the mechanism of how CSB regulates R-loops before TCR.

Here, we demonstrate that R-loops are significantly induced at transcription start site and gene body when CSB is depleted. Moreover, a new type of R-loops associated with poly T tract on the non-template strand are very sensitive to CSB depletion, indicating that CSB are recruited to the poly T loci during transcription. Importantly, we found that GRO-seq signals decreased sharply at the poly T tract, showing that such poly T associated R-loops are induced due to pol II pausing which was also confirmed by in vitro transcription assay, These findings suggest that CSB not only functions as a TRC initiation factor, but also a global R-loop regulator. It also provides new insights in elucidate the underlying mechanism of the Cockayne syndrome.

3.2 The R-loop landscape in response to CSB depletion

Because CSB is the key initiation factor in TCR[Xu et al., 2017a], previous studies all focused on examining the interaction of CSB with the accumulated R-loop in cis of DNA dam-

ages (DSB or SSB) provoked by various DNA-damaging agents, such as UV irradiation, targeted reactive oxygen species and nucleases, in different model systems[Jang et al., 2020, Teng et al., 2018, D'Alessandro et al., 2018]. However, CSB should also be able to interact with transcriptional coupled R-loops, as it is one of the first proteins in sensing the presence of R-loops during TCR. Therefore, we didn't apply any DNA-damaging treatment in our system. To uncover the mechanism of how CSB regulates R-loops, we took advantage of the strategy developed in our lab, R-ChIP to map R-loops genome-wide by using the catalytically inactive RNase H1, which greatly increases the resolution of R-loop signals compared to the S9.6 based R-loop profiling method. In HEK293T, R-loops were profiled in vivo under normal condition and knockdown of CSB condition by siRNA.

If CSB is recruited to TCR induced R-loops exclusively, we would expect R-loop landscape remains the same when there is no DNA lesion introduced in both conditations. However, we found that CSB knockdown dramatically reform the R-loop landscape as illustrated by both the specific gene examples and genome-wide meta-analysis. Specifically, we compared the R-loop profiles in two conditions in the following aspects: peak number, peak length distribution, peak shape and their genomic location distributions (Figure 3.1, Figure 3.2).

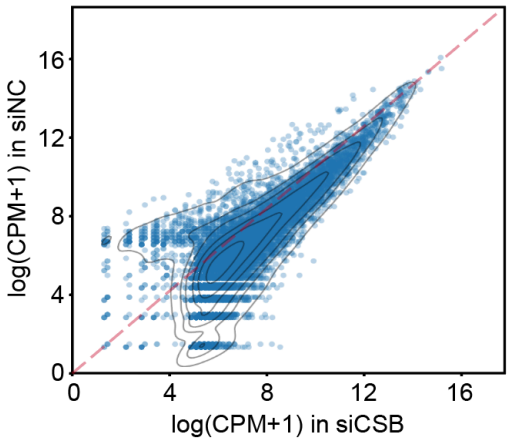

Figure 3.1: **Scatter plot of R-ChIP.** Scatter plots showing R-loop signal profiled by R-ChIP in CSB-nondepleted cells (siNC; Y-axis) versus CSB-depleted cells (siCSB; X-axis). Each dot represents a R-loop; Grey lines on the plots represent the density contour of dots. The scale for both the X-axis and the Y-axis is log(CPM+1); CPM is count per million.

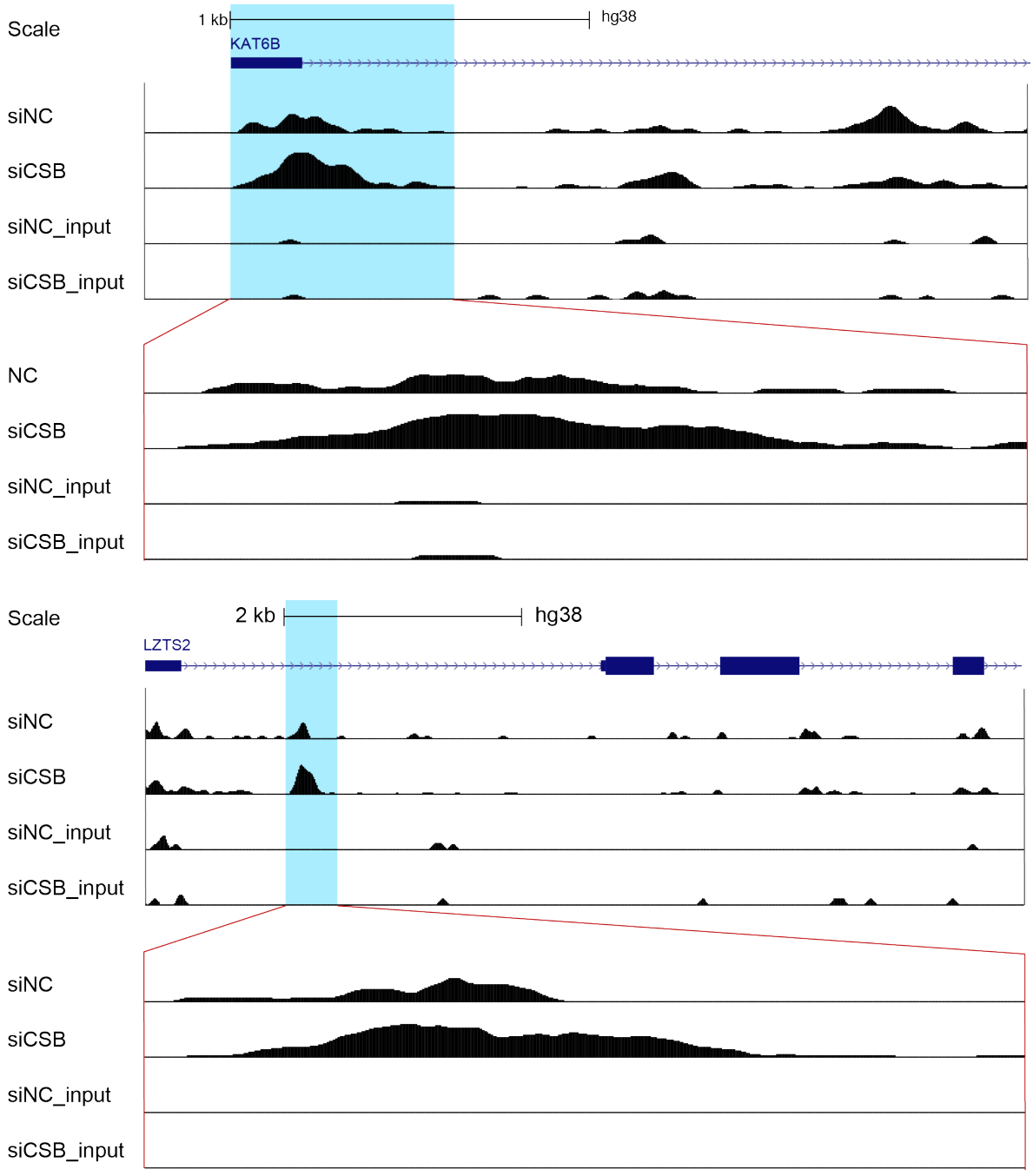

Figure 3.2: Genome browser tracks showing a region of R-loop signals in R-ChIP in CSB-depleted and nondepleted cells and their input signals. Gene annotations are shown on top of the tracks.

Remarkably, we found that the number of R-loops is more than doubledin the CSB-knockdown cells (7750 R-loops) compared with control cells (2775 R-loops)(Figure 3.3). The length distribution of CSB-knockdown cell (median 180 bp) also increased around 30 bp compared with the control cell (median 210 bp) (Figure 3.3). Moreover, in the CSB-depleted cells,

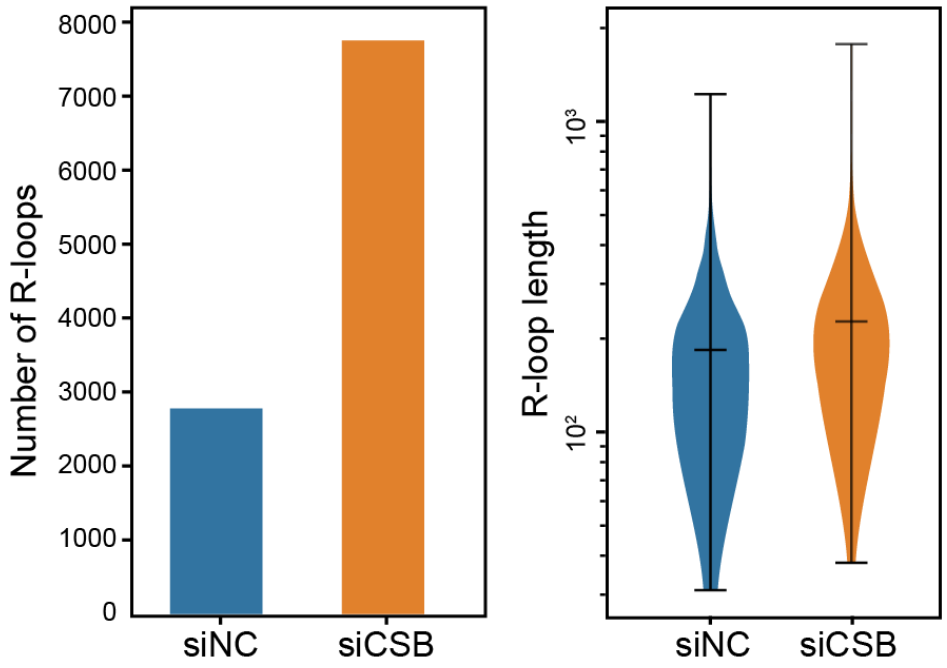

Figure 3.3: **R-loop number and length distribution.** (Left) Bar plot showing the number of R-loops under siNC and siCSB conditions. (Right) Violin plot showing the R-loop peak length distribution under siNC and siCSB conditions.

.

the R-loop intensity is 1.5-fold to the R-loop intensity of control cell on average and the R-loop center of the CSB-knockdown cell slightly shift to downstream of the transcription direction (Figure 3.4). We further checked the R-loop genomic location distribution and plot the Venn diagram of R-loops mapped at TSS, gene body, TTS and intergenic regions in CSB knockdown cell vs control knockdown cell. In both conditions, ~45% R-loops formed at TSS region, ~43% R-loops formed at gene body region, ~2% R-loops formed at intergenic region and ~10% R-

loops formed at intergenic region (Figure 3.5). While most of the newly generated R-loops are observed in TSS (71.3% TSS R-loops in siCSB are gained) and gene body (73.5% genebody R-loops in siCSB are gained) region. Taken together, CSB knockdown induced more R-loops, longer R-loops and larger R-loops across the whole genome uniformly, suggesting its important role in R-loop regulation.

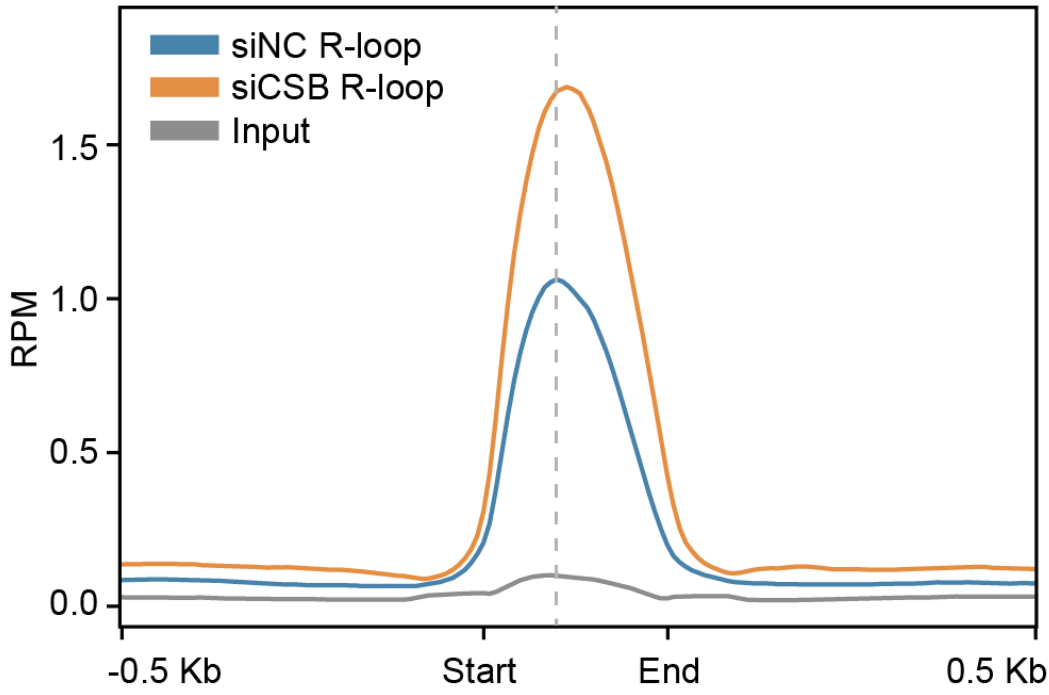

Figure 3.4: **Meta plot of R-loops in siNC and siCSB.** Meta signal representations of the mean R-ChIP-seq signal under siNC (Blue line) and siCSB (orange line) conditions. signal are centered on the R-loop summit, additional 0.5 kb surrounding each summit is also shown.

.

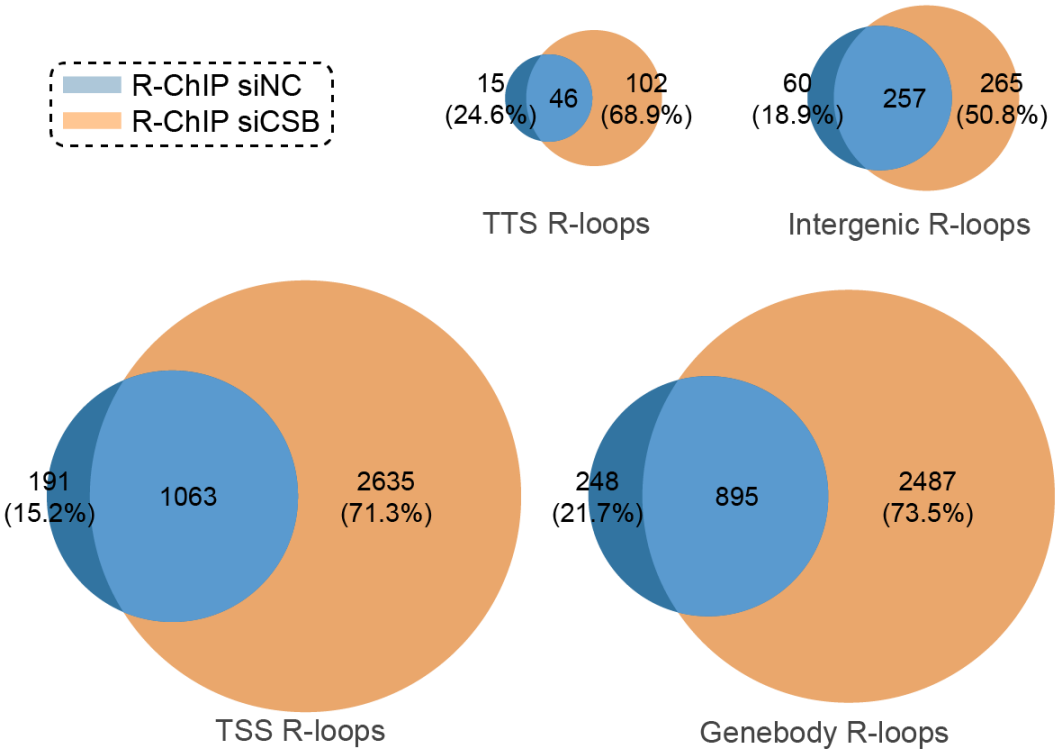

Figure 3.5: **Venn digrams of R-loops at multiple genomic locations.** Venn diagrams depicting the overlap between R-loops under siNC and siCSB conditions in TSS, Genebody, TTS and Intergenic regions.

.

3.3 R-loops formation at poly T tract in gene body

3.3.1 Poly T feature in non-template strand

Evidence shows that CSB is not only responsible for the initiation of TCR, but also could stimulate the rate of elongation on an undamaged template [Selby and Sancar, 1997], which led us to examine whether the R-loops induced by CSB knockdown are associated with new features, specifically, motif features and genomic location preference. Because most of R-loops detected in control cells (siNC) overlapped with that detected in CSB knockdown (siCSB) cells. Motif

analysis was performed on the common 2261 R-loops shared by siNC and siCSB, and on the gained 5489 R-loops identified only in CSB knockdown cells. As R-ChIP is a strand specific method to detect DNA/RNA hybrids, we thus perfromed the strand-specific motif analysisuse using the ssDNA sequence of R-loops as the input, which is also the non-template strand sequence of the transcribed RNA. We found only the GC rich motifs showed significant enrichment on both common and gained sets of R-loops. No new motif features emerged in the gained R-loops when we analyzed the motif on the 100 bp sequence centered at the summit of identified R-loops.

Because CSB tended to be recruited to pol II when ”accidents”, such as DNA lesions, happened during transcription, we decided to examine potential sequence features the proximity of induced R-loops. Thus, we further divided each R-loop into two segments (Figure 3.6), the head and the tail sections corresponding to the upstream and downstream sequences in reference of the R-loop summit position. Similar to the motif features identified at the R-loop summit, both the common and gained sets of R-loops enriched the GC-rich motif at the head section of R-loops. However, a consecutive T tract is enriched significantly at the tail section of R-loops only in the gained R-loops set with CSB knockdown. High GC-skew (quantified by (G-C)/(G+G) in a segment) on the ssDNA of R-loops has been long thought as a key feature of R-loops, because of the potential formation of secondary structure, such as G4, on the looped-out ssDNA within R-loops. Moreover, high AT-skew (quantified by (A-T)/(A+T) in a segment) in the ssDNA has also been detected in the R-loops formed in the yeast genome. This is the first time that a T-rich tract on the non-tempate strand is linked to R-loops in a CSB-dependent manner.

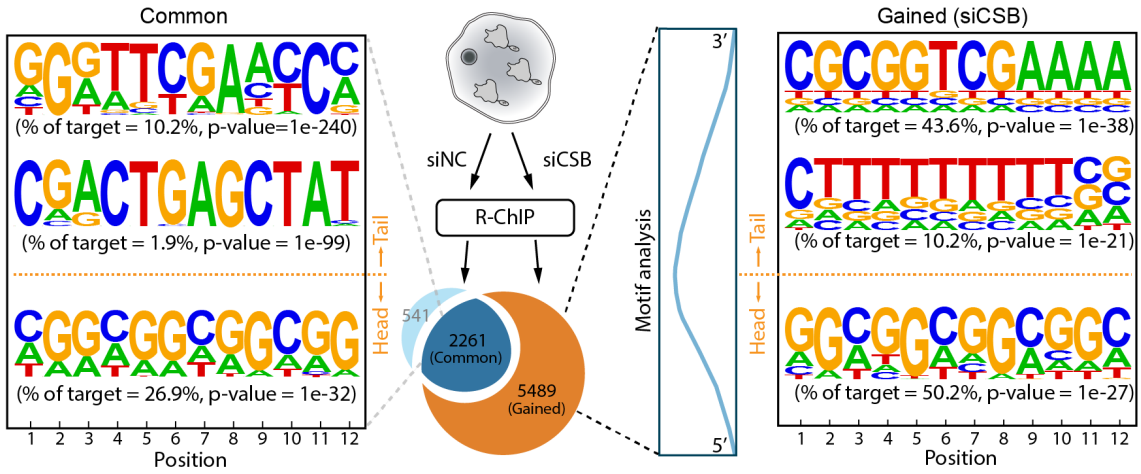

Figure 3.6: **Motif analysis on R-loops.** Table on the left depicting the common R-loops (shared by siNC and siCSB) motifs enriched at the head and the tail within whole R-loop regions and associated p values. Table on the right depicting the unique R-loops (induced by siCSB) motifs enriched at the head and the tail within whole R-loop regions and associated p values.

.

3.3.2 Collaborative contribution of GC skew and Poly T to R-loop induction

To assess whether the poly T feature at the tail of gained R-loops is able to induce R-loops independently, we further checked GC-skew and T percentage across all R-loops to quantify and compare these two features in the siNC and siCSB conditions. If merely poly T tract could induce R-loops, it means that such kind of R-loops could form without the requirement of the high GC-skew in the head section. Therefore, lower GC skew and higher T percentage are expected to be observed under siCSB condition versus siNC condition. Indeed, as shown in Figure 3.7, the T percentage is higher in siCSB than that in control started from the R-loop summit to the tail section, although the average T frequency across the R-loop region is below 25% (if we assume the equal frequency of the A, T, G, C). For the GC-skew, it hit 0.3 (means 65% G over 35% C in GC portion) ahead of the R-loop summit which is constant with our previous observation on the

general R-loop features. These results indicate that, the CSB induced R-loops associated with poly T still featured with a high GC-skew in the head section.

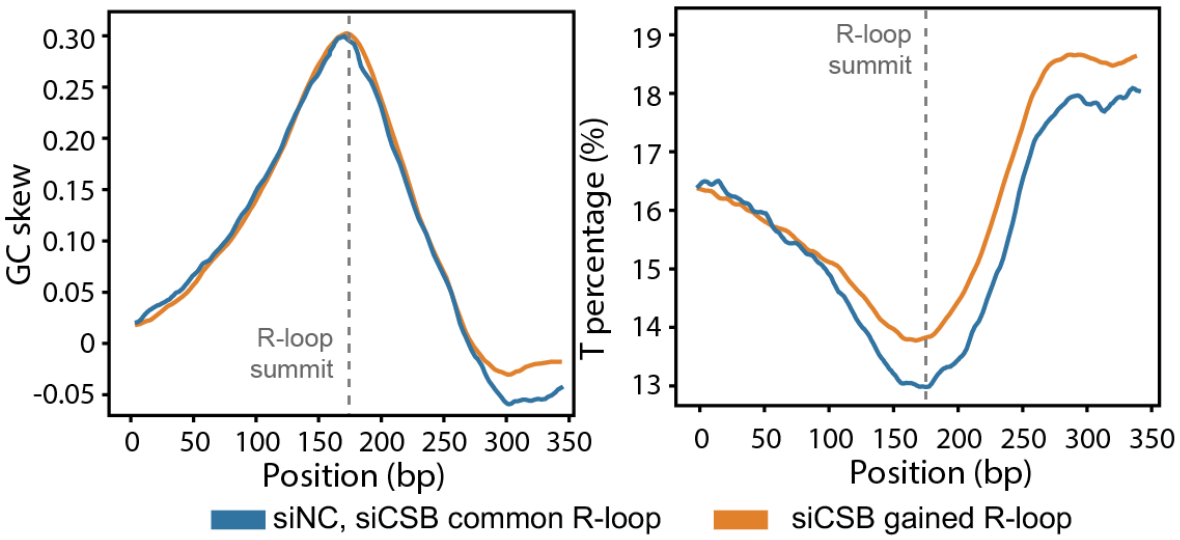

Figure 3.7: **GC skew analyses and average T percentage on R-loops.** (Left) The average of the GC-skew (G - C/G + C) in the ssDNA of R-loops in siNC and siCSB are represented by blue and orange lines, respectively. (Right) The average of the T percent in the ssDNA of R-loops in siNC and siCSB are represented by blue and orange lines, respectively. All were calculated by using a window size of 50 bp and a step size of 1 bp.

.

Then, we isolated the R-loops associated with poly T under both conditions. Consistent with the result of the motif analysis, there are ~750 poly T associated R-loops identified in siCSB cells, while ~150 identified in siNC cells (Figure 3.8). Importantly, we also found that, in siCSB, 68% of the poly T R-loops located in the gene body region. While, in siNC, 45% of the poly T R-loops are detected in gene body region, which means that most of the poly T R-loops gained by CSB knockdown fell in the gene body region. This observation is consistent with the fact that CSB has been shown to function together with pol II as an elongation factor. It suggests that these poly T R-loops are directly induced by the absence of CSB during transcription.

In order to further decipher how these two features collaboratively contribute to R-loop

formation, we calculated the GC-skew in the poly T R-loop subset in siCSB condition. Comparison between all loops detected in siCSB treated cells vs those associated with T tracts revealed decreased GC-skew value from 0.3 to 0.2. As illustrated in the Figure 3.9, we plotted the average GC-skew across the R-loop region aligned with the poly T tract in the tail section, where the highest GC-skew hit 0.2 at the head section. Importantly, we also found that GC-skew is negatively correlated with the length of the poly T track, the longer the T tract exists, the lower the GC-skew observed in the R-loops head section. It thus indicates that, during transcription, CSB may act as a general regulator at poly T tract to prevent or resolve the formation of R-loops, and the underlying mechanism must be relevant with the length of the repetitive T sequence.

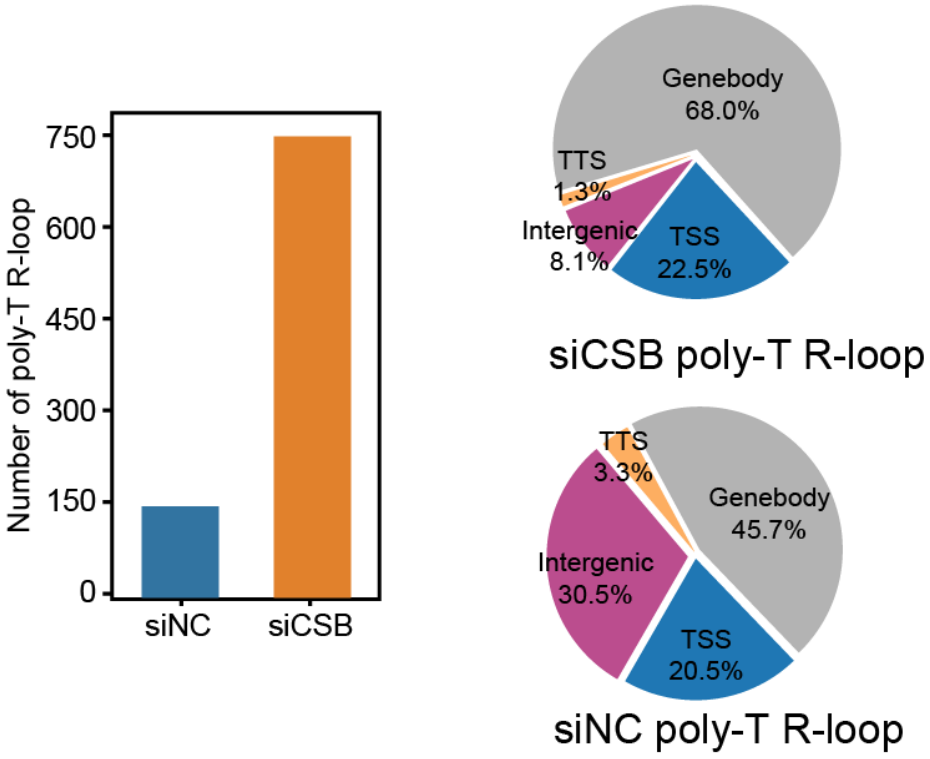

Figure 3.8: **Poly T R-loops number and their genomic location distribution.** (Left) Bar plot showing the number of R-loops associated with poly T tract under siNC and siCSB conditions. (Right) Pie plots showing the Poly T R-loops genomic location distribution under siNC and siCSB conditions.

.

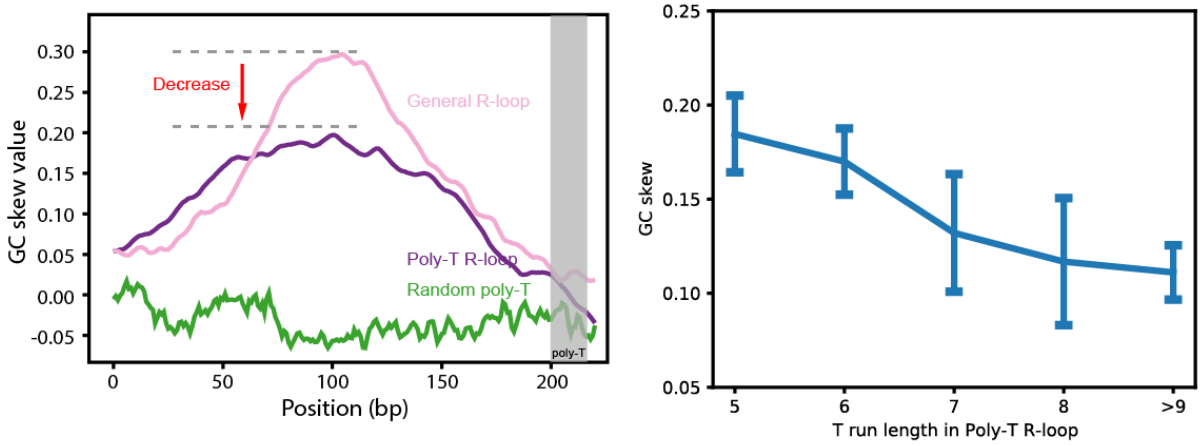

Figure 3.9: **GC skew analyses on R-loops associated with poly T.** (Left) The average of the GC-skew (G - C/G + C) in the ssDNA of poly T R-loops in siCSB, the normal R-loops without poly T and the random poly T sequence are represented by purple, pink and green lines, respectively. All were calculated by using a window size of 50 bp and a step size of 1 bp.(Right) The average GC-skew in the ssDNA of poly T R-loops, where poly T R-loops are splited by the length of the poly T tract patterns.

.

3.4 CBS-induced R-loops through the direct interaction with RNA Pol II

3.4.1 CSB colocalized with Pol II at R-loops during transcription elongation

To investigate possible role of CSB in R-loop formation, CSB binding was measured genome-wide by ChIP-seq in the stable cell line overexpressing the CSB-3XFlag to achieve high Immunoprecipitation (IP) efficiency. We found that the majority of CSB binding sites fell in gene body region (61%), while 30% of them distributed at TSS region, with 6.2% of them in TTS and only 2.4% in intergenic region, suggesting that, similar as R-loop, CSB binding events are tightly linked to active transcription. Given the fact that CSB, as an elongation factor, forms complex with Pol II during TCR initiation, we generated the heatmap of CSB binding and R-loop intensity at Pol II pausing sites measured by Pol II ChIP-seq in gene body region (Figure 3.10). As expected, CSB ChIP-seq signals highly correlated with Pol II binding sites. Due to the fact that more R-loops were induced when CSB is depleted, in CSB overexpressing cell, CSB binding dynamics around R-loop region may be hard to detect. However, we found that most induced R-loops in siCSB treated cells were coincident with CSB binding events in the genome.

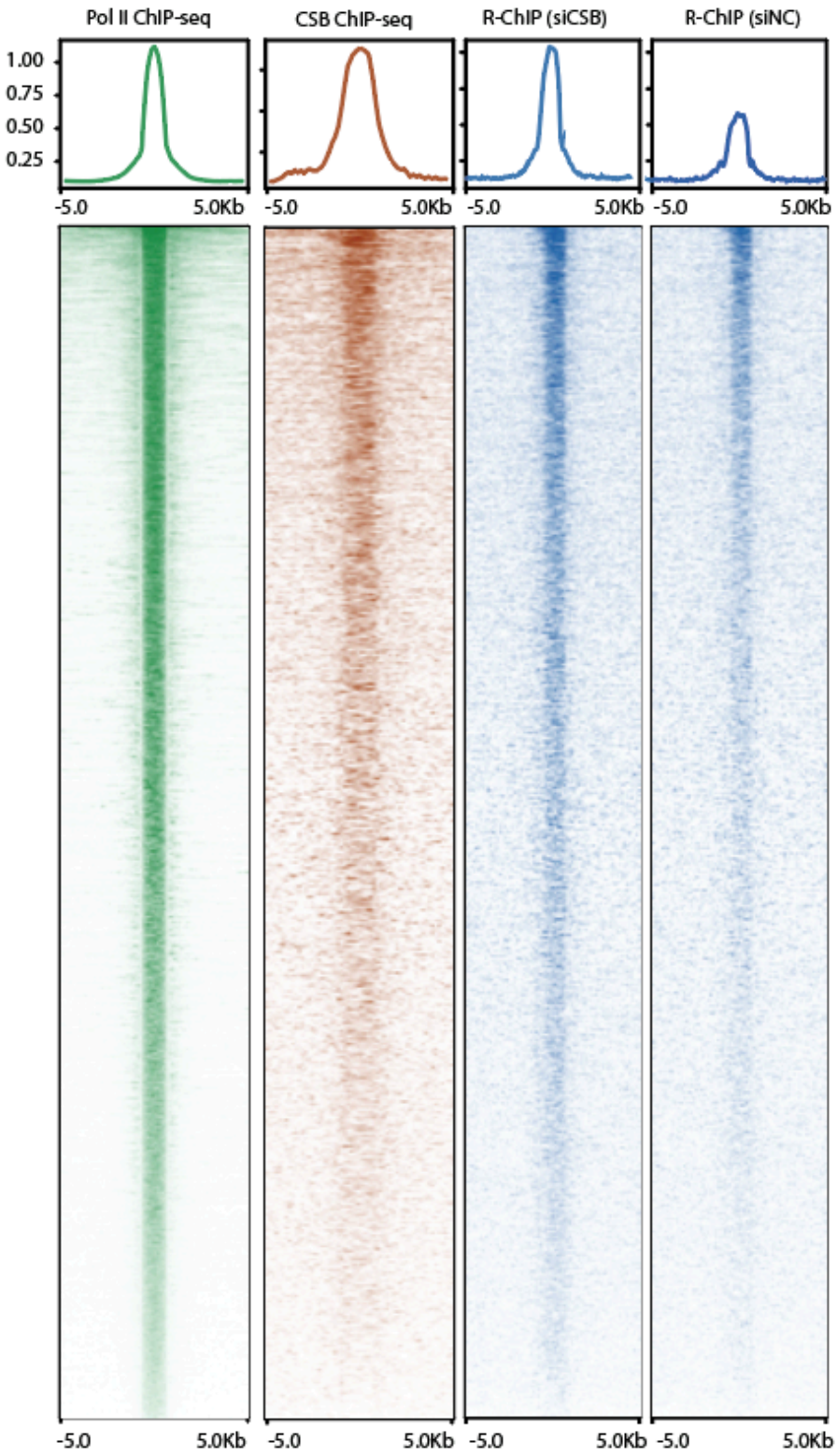

Figure 3.10: Heatmap showing the binding pattern of Pol II, CSB and R-loops pattern before and after CSB depletion. Color scale reflects the intensity of the signals.

We next turned our attention to the poly T tract in the gained R-loop by CSB depletion, specifically, by inspecting the occupancy of Pol II and CSB. Not surprisingly, CSB is not only enriched at the overall induced R-loop region, but also shows a finer binding pattern around the poly T tract within R-loops. Besides, we also detected Pol II enrichment. Meta-analyses (Figure 3.12) were also performed on poly T tracts identified from 750 poly T R-loops described before in siCSB. In line with the previous results, we found that Pol II binding started to decrease from the poly T sites, which may potentially provide a pausing platform waiting for the recruitment of CSB to form a functional complex. Importantly, although the absolute CSB ChIP-seq signal is not too high due to the attempting to capture a very dynamic biological process, a sharp enrichment could be still detected at poly T tracts. According to the R-loop profile of siNC and siCSB plotted in the meta analyses, R-loop intensity increased more than 80% right ahead of the poly T tract. Strikingly, the GRO-seq signal fell down like a water fall at the poly T track, which additionally validates the possible pausing of transcription at these poly T tracks within CSB depletion induced R-loops.

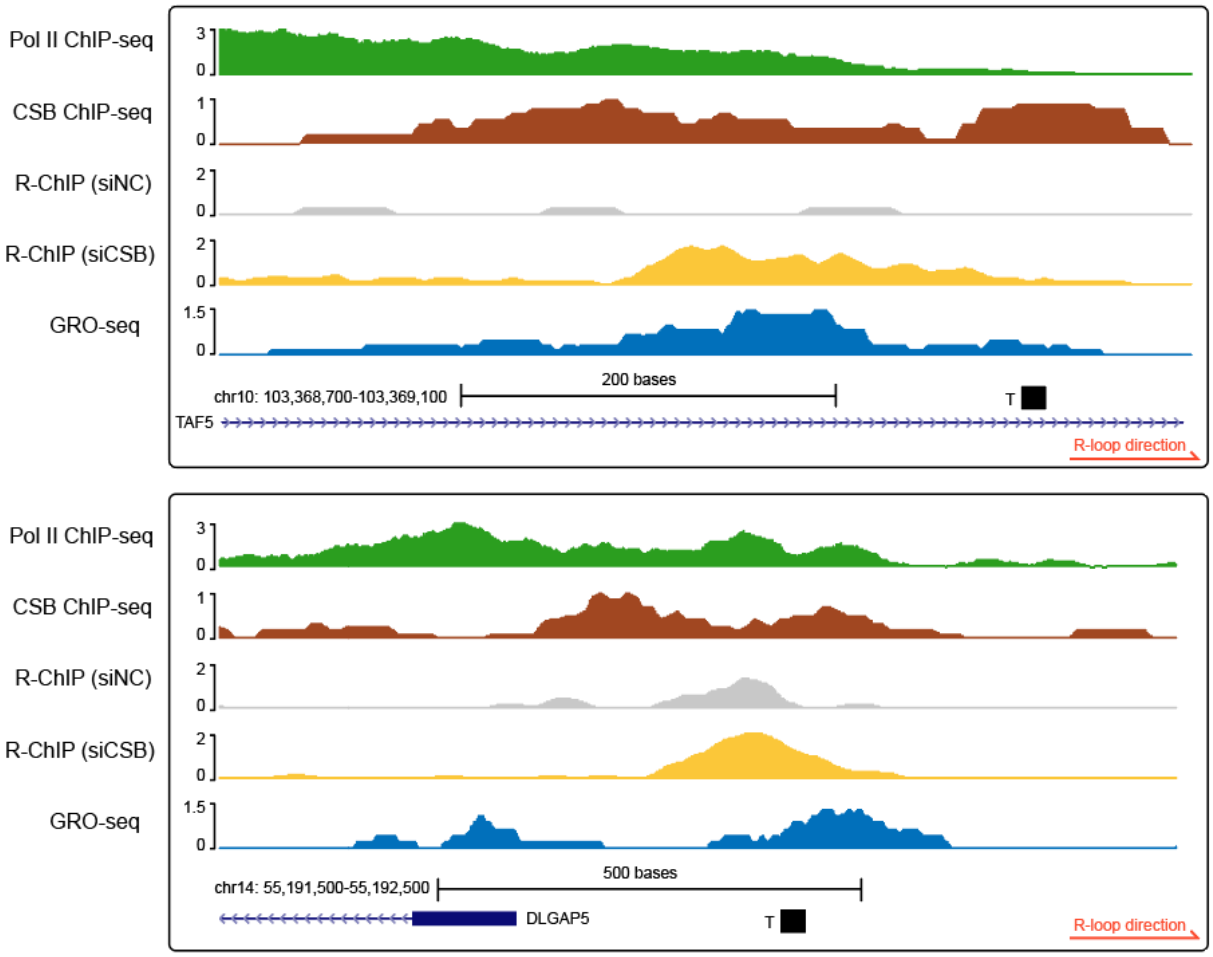

Figure 3.11: R-loop located in TAF5 gene body was induced by CSB depletion. (Top) R-ChIP signal detected in siCSB and siNC shows that R-loop is induced at gene body with CSB depletion (Middle) CSB binding is observed at the poly T tract located at the tail of the R-loop. (Bottom) Decrease of the GRO-seq signal is observed ahead of the poly T tract.

.

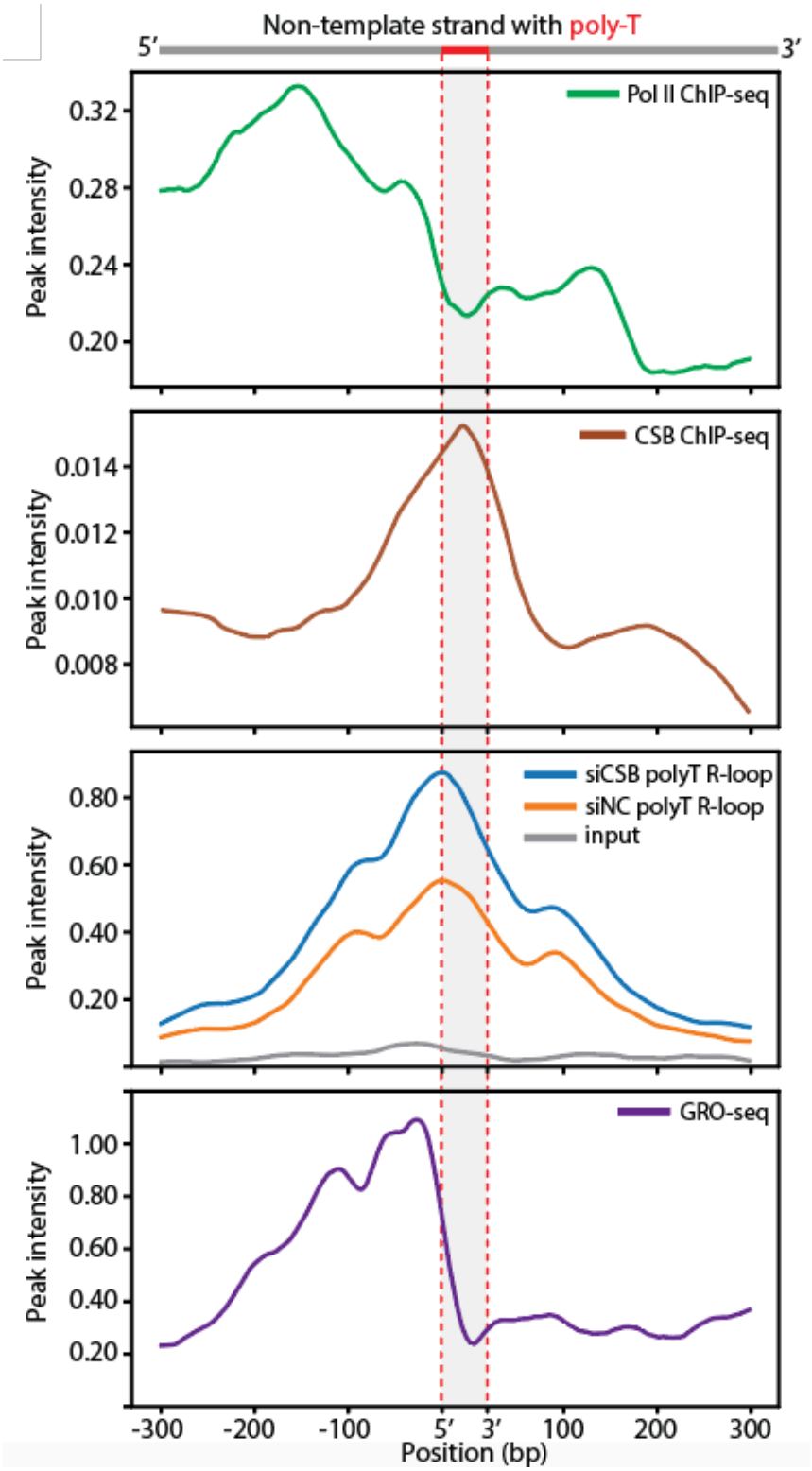

Figure 3.12: **Meta profile at poly T R-loop.** Meta signal representations of the mean R-ChIP-seq signal under siNC (Blue line) and siCSB (orange line) conditions. Meta Gro-seq signal is shown by the purple dash line. Signals are centered on the poly T tract at the tail of poly T R-loops, additional 0.3 kb surrounding each poly T tract is also shown.

Altogether, these results show that Pol II tends to pause at poly T tract during transcription, and CSB may form complex with the paused Pol II to prevent the potential transcription abortion. Pol II may fail to jump out of the poly tract in the CSB depleted cells, thus, leaving the unfinished transcript and jam the coming transcription activity to induce transient R-loops form ahead of poly T tracts. Pol II may back tracked from the poly T tract in CSB depleted cells, thus leaving nascent RNA to form R-loops in the front of the polymerase. Furthermore, apart from the coincidence of CSB and pol II detected in the gene body, CSB ChIP-seq signal also enriched at the Pol II binding sites in TSS region, suggesting a more general role of CSB in Pol II pause release independent of polyT tracts.

3.5 Mechanism underlying CSB-regulated R-loop

3.5.1 CSB facilitated transcription elongation

To further validate our hypothesis and reveal the underlying mechanism, we reconstituted an in vitro transcription elongation system using the template sequence acquired from our in vivo sequencing data, based on the purified Pol II components and the yeast CSB ortholog, Rad26. Because we found a sharp fall of both Pol II ChIP-seq and GRO-seq signals at the poly T tract, first, we would like to test whether Pol II paused on the poly T tract within the identified poly T R-loops.

In the in vitro transcription system (Figure 3.13), radioactively-labeled RNA, template strand DNA and Pol II were assembled first. Then, biotin labeled non-template strand was added to from the upstream elongation complex. The assembled elongation complex was purified and

isolated by Streptavidin magnetic beads, forming a bead associated elongation complex which is ready to ligate with downstream DNA template. On the other hand, poly T tract which is selected from the poly T R-loop subset in siCSB, has been amplified from the genomic DNA to serve as the downstream template. Finally, the upstream sequence associated elongation complex and the downstream poly T template were ligated, providing an ideal platform for us to observe how Pol II behave at the selected poly T tracts.

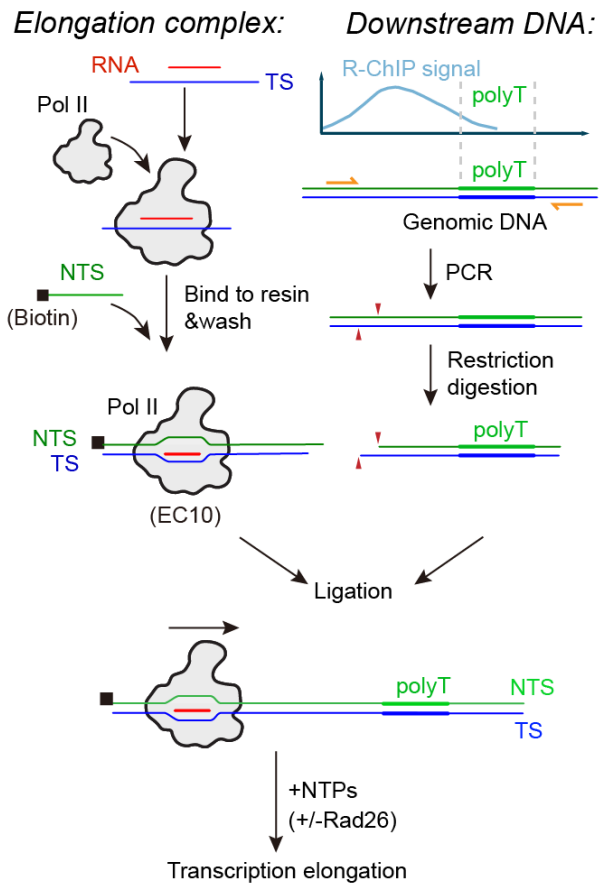

Figure 3.13: A scheme of the experimental setup to study the function of CSB in Pol II transcription on the poly T R-loop. A Pol II EC segment and a poly T containing DNA fragment (R-loop identified from R-ChIP) were reconstituted individually and ligated by T4 ligase. A biotin label is shown as a black square.

The full run-off RNA would be 220 bp when Pol II successfully carried out transcription without falling off halfway. We would expect to observe shorter run-off products if Pol II was trapped at poly T tract. Also, we performed the time course transcription assay with collecting the produced transcript at four time points, 30s, 1 min, 3 min and 10 min. As expected, compared to the control template where has no poly T tract, we observed two intermediate run-off products around the poly T tract as labeled in Figure 3.14, representing aborted transcription at the Poly T tract with purified Pol II without assistance of other co-factors, such as CSB. We also found that more full run-off transcripts were transcribed as we increased the incubation time, which means Pol II itself could have chance to jump out of the poly T trap without the help of other factors in our in vitro system.

We next tested how CSB functions on the paused Pol II at the poly T tract template during elongation. We found that, in the presence of CSB, there is a significant increase of the full length run-off transcript over the reaction time period compared to the buffer only condition, suggesting that Pol II escaped from the pausing state and bypassed the poly T trap efficiently. Each in vitro transcription assay has been repeated three times and the intensity change of each run off product was shown in the bar plot as well. Overall, our results further validate that Pol II paused at that poly T tract and CSB could help paused Pol II overcome the poly T trap and thus promoted the transcription efficiency.

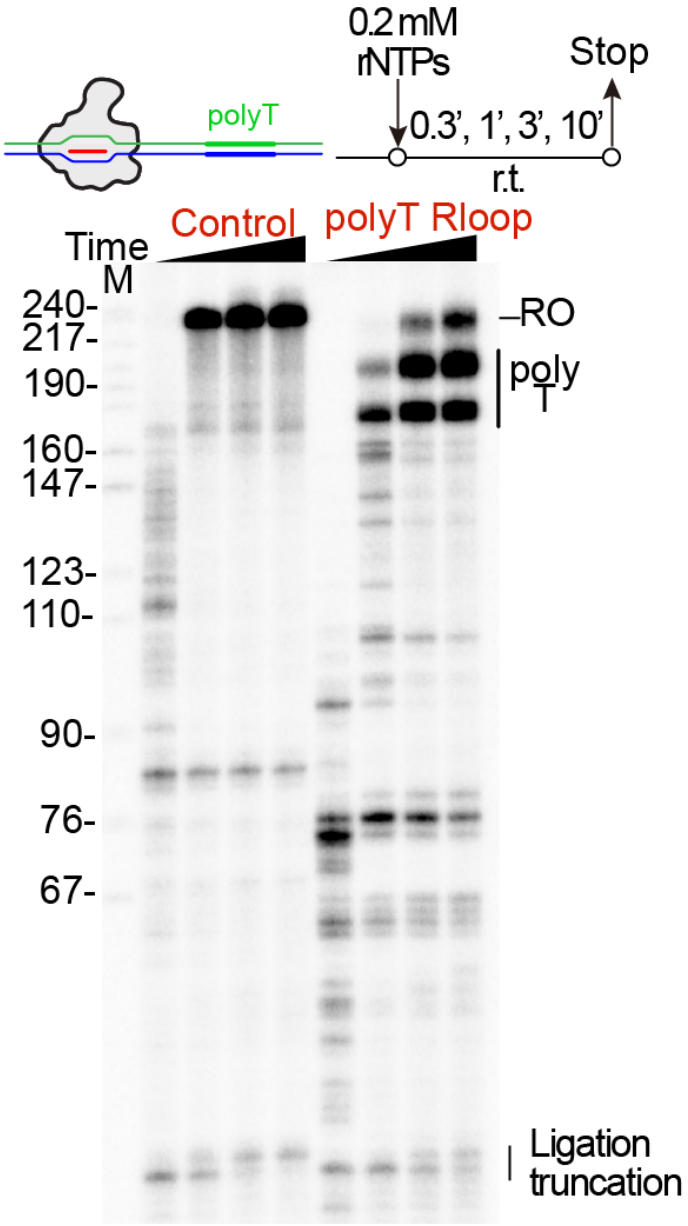

Figure 3.14: **PolII paused at poly T tract.** (Top) Schematic of the transcription reaction. (Bottom) Mapping of pausing sites during Pol II transcription on a control sequence without poly T tract (Left) or on a poly T R-loop template (Right). Arrest sites are indicated on the right. The ligation truncation corresponds to the transcripts from reconstituted ECs that did not ligate to the downstream poly T R-loop template.

.

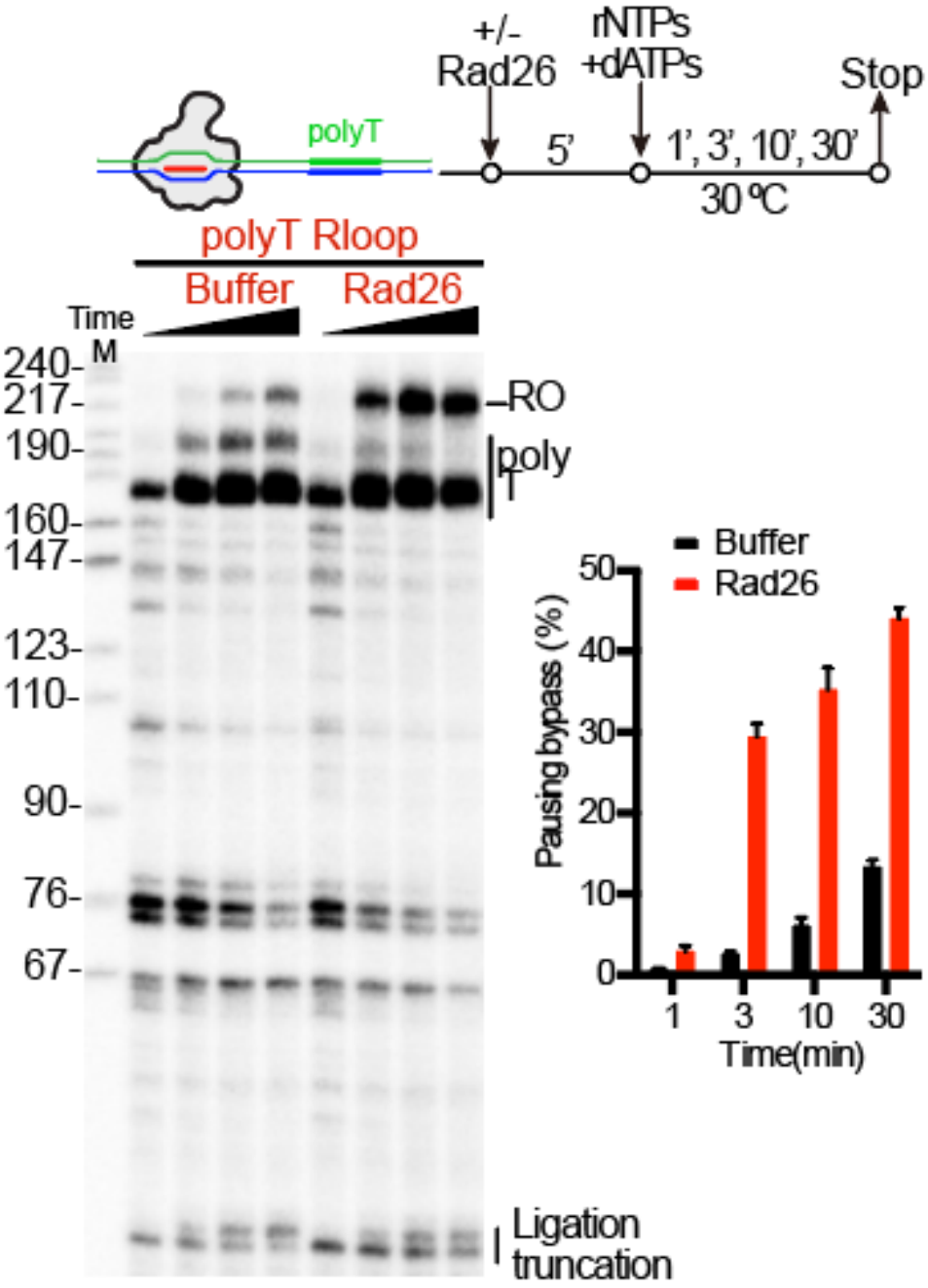

Figure 3.15: **Contribution of Rad26 on promoting Pol II bypass poly T tract.** (Top) Schematic of the transcription reaction. (Bottom) Mapping of pausing sites during Pol II transcription of a poly T R-loop template in the absence of Rad26 (Left) or in the presence of Rad26. Arrest sites are indicated on the right. The ligation truncation corresponds to the transcripts from reconstituted ECs that did not ligate to the downstream poly T R-loop template.Bar plot shows the quantitation of the native PAGE assay.The experiments were performed three times and shown as means with standard deviation error bars.

.

3.5.2 Poly T-induced R-loops during elongation

To verify that whether the paused Pol II could induce R-loop on selected poly T tract, we next measured the R-loop level through RNase H treatment using the in vitro transcription assay. We first set up the previously described transcription reaction for 30 min. Then, RNase H was used to assess the formation of R-loops, the digested transcripts were collected at 0 min, 1 min, 2 min and 5 min. If R-loops formed ahead of poly T tract as we observed from the R-ChIP signal, it would be removed and digested by RNase H. As expected, both the rull run-off product and the paused band signal decreased significantly under the RNase H treatment. Besides, we also captured the digested shorter fragments as indicated by the blue line in Figure 3.16, the intensity of which increased with the longer RNase H treatment. In contrast, none of the digested bands were detected in transcription assay using control template contains no poly T tract. Our results suggested even under the in vitro transcription condition, poly T sequence is able to induce R-loops recognized by RNase H.

3.5.3 CSB-mediated R-loops resolution by pushing Pol II forward

Although such poly T R-loops are induced by CSB depletion, whether CSB can directly resolve R-loops remains unknown. Based on our previous results, CSB is very likely to resolve R-loop by pushing Pol II forward [Xu et al., 2017a]. To test out the final piece of the puzzle, we measured the R-loop amount of the poly T R-loop in the presence or absence of CSB. As we described previously, once we set up the in vitro transcription elongation, we treated it with the wildtype Rad26 and the ATPase-deficient mutant of Rad26 who abolishes its DNA translocase activity in promoting Pol II transcription over DNA barriers [Xu et al., 2017a]. Both the mutant

and the buffer condition served as the negative control to the wild type Rad26 treatment. With the 40 min CSB treatment, RNase H1 was added to the system to quantify the relative R-loop amount. As shown in Figure 3.17, with the WT CSB, the intensity of full run-off band increased significantly. While the amount of the pausing transcripts reduced a lot, indicating that Pol II was no longer arrested at poly T in the presence of CSB. Compared to negative control, the RNase H digested short transcript also disappeared with the help of CSB, suggesting that CSB resolves the R-loop induced by the poly T tract.

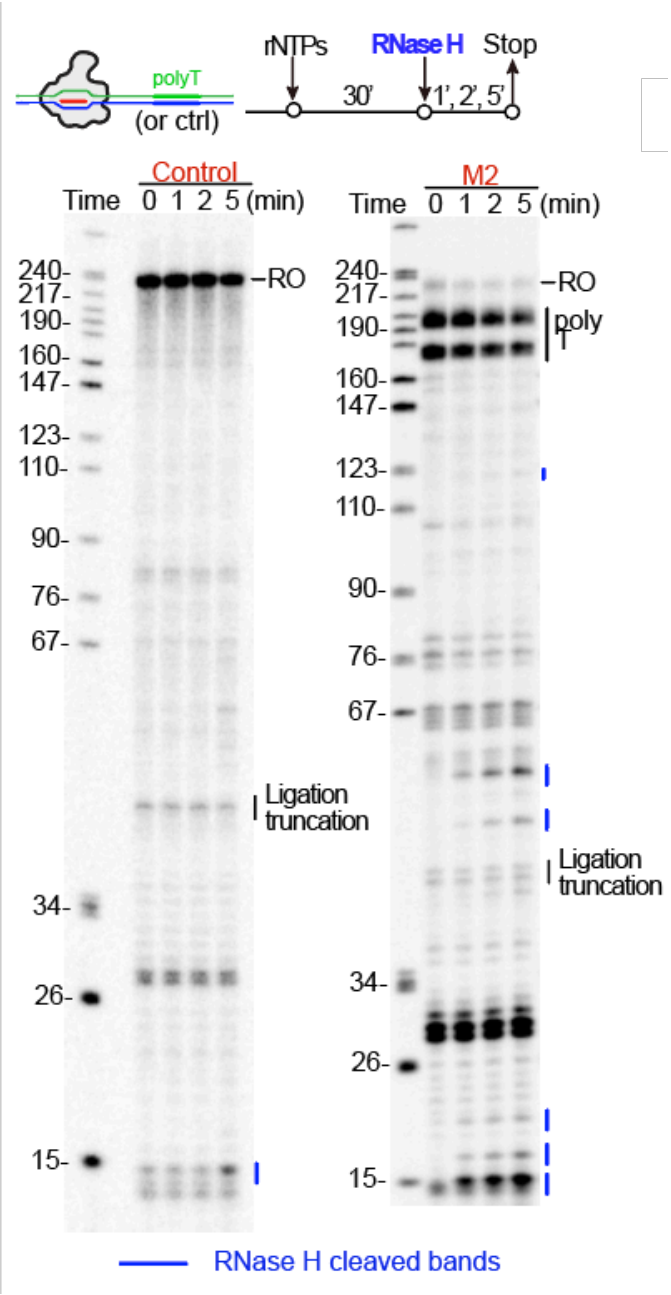

Figure 3.16: **R-loop caused by poly T tract during transcription.** (Top) Schematic of the transcription reaction. (Bottom) High molecular weight transcript from poly T on NTS DNA template was cleaved by RNase H, while the on the control template without poly T tract, the run off trancripts were not digested by the RNase H. The blue lines indicate RNase H digested transcript. The black lines indicate the RNA transcript from template with incomplete ligation (ligation truncation).

.

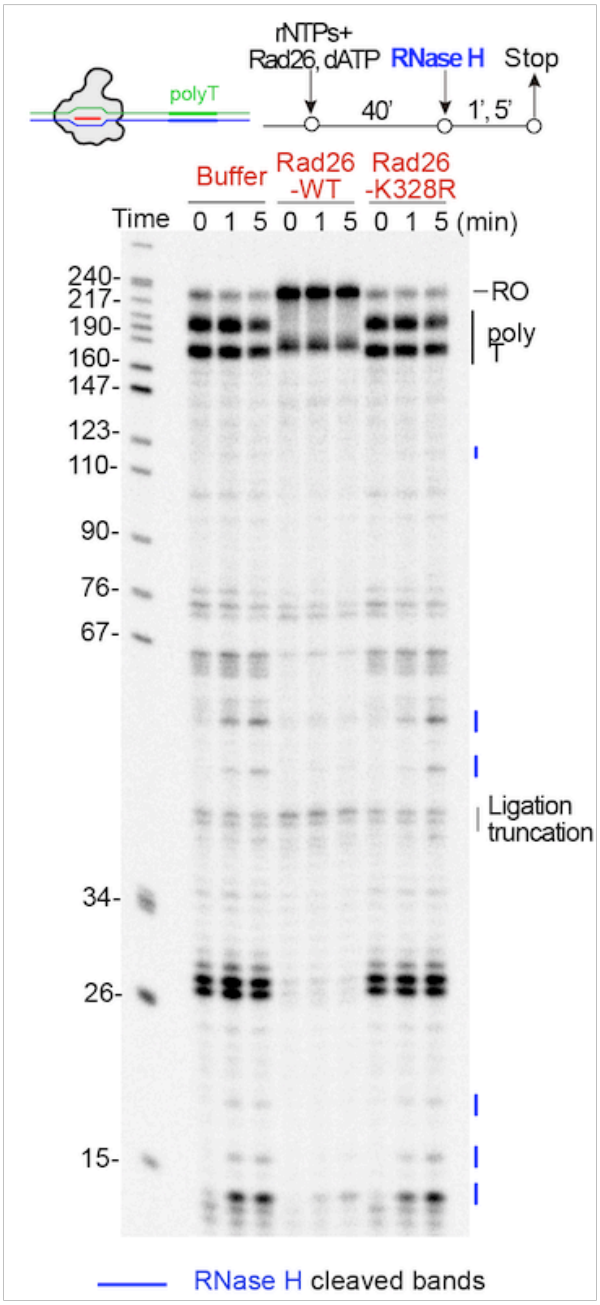

Figure 3.17: **Contribution of Rad26 on R-loop resolution.** (Top) Schematic of the transcription reaction. (Bottom) RNase H assay on Poly T tract during Pol II transcription in the absence of Rad26 (Left) or in the presence of either WT (Middle) or an ATPase-deficient mutant (Right) of Rad26.

.

3.6 Role of R-loop in Cockayne Syndrome

3.6.1 Contribution of mutant CSB to Cockayne Syndrome

Cockayne Syndrome is a rare disorder that is caused by one of the genes (CSA and CSB) having mutations. Those who have this disease have an appearance of short stature and premature aging. They suffer through many symptoms, along with growth failure and increased sensitivity to the sun. The underlying disorder is a defect in transcription coupled DNA repair. The TCR pathway is a process of removing DNA lesion in a transcription-dependent manner. During elongation there will always be some road blockers to impede Pol II moving forward to induce pol II pausing, such as mismatchs and other types of DNA lesions. Once Pol II hits the road blocker, it has chance to stall there. CSB, as the elongation factor, is then recruited to the paused pol II and release signals for other transcription coupled Repair proteins to fix the DNA damage. Pol II may wait during the fixation process, it is very likely to form R-loops. As poly T R-loops are easily induced when CSB is depleted, it may play important roles in the development of Cockayne Syndrome. Based on this hypothesis, we woule like to analyze the poly T tract distribution in the CSB sensitive genes whose expression level are dependent on the CSB expression.

3.6.2 Enriched poly T tracts in CSB sensitive genes

To investigate how the poly T tract functions in Cockayne Syndrome, we perform the differential gene expression analysis on two sets of RNA-seq experiments to identify CSB regulated genes. The frist set of RNA-seq was performed on U20S cells before and after CSB depletion. The second set of RNA-seq was performed on CSB deficient Cockayne Syndrome patient cells

and the those cells under rescuing condition, with CSB overexpressed. In both sets, more genes were upregulated when CSB expressed normally compared with the number of downregulated genes. We then split the genes into three polulations, CSB downregulated genes, CSB upregulated genes and CSB insensitive genes, based on the fold change 2 and p-value 0.01 thresholds. We found that genes that were significantly down regulated in response to CSB depletion contain more polyT tracts per gene compared to the other two gene populations, suggesting a positive correlatio between the presence of polyT tract and CSB dependency. Furthermore, we inspect the gene length distribution of the three gene sets, as the high poly T tract containing feature may related with the gene length. Not surprisingly, CSB upregulated genes are significantly longer than the genes in the other two sets as shown in Figure 3.18. Besides, we also found that CSB upregulated genes contains slighly more poly T tract per kilo base in U2OS cells, while they had almost the same poly T tract frequencies compared with the other two gene sets (Figure 3.19). These results indicate that a set of long genes expression are very sensitive to the CSB depletion, the longer the gene is, the higher tendency for the gene to harbor more polyT tract, which would be more prone to R-loop formation in CSB mutant cells. These excess poly T-induced R-loops may contribute to Cockayne Syndrome in response to CSB deficiency due to increased transcription failure on relative long genes, which is more prevalently expressed in the central nervous system [Gabel et al., 2015].

3.7 Acknowledgements

Chapter 3, in full is currently being prepared for submission for publication of the material. X. Zhang, J. Xu, Y. Hao, D. Wang, X.-D. Fu, “CSB Resolves R-loops by resuming stalled Pol II

at poly T tract". The dissertation author was the primary investigator and author of this paper.

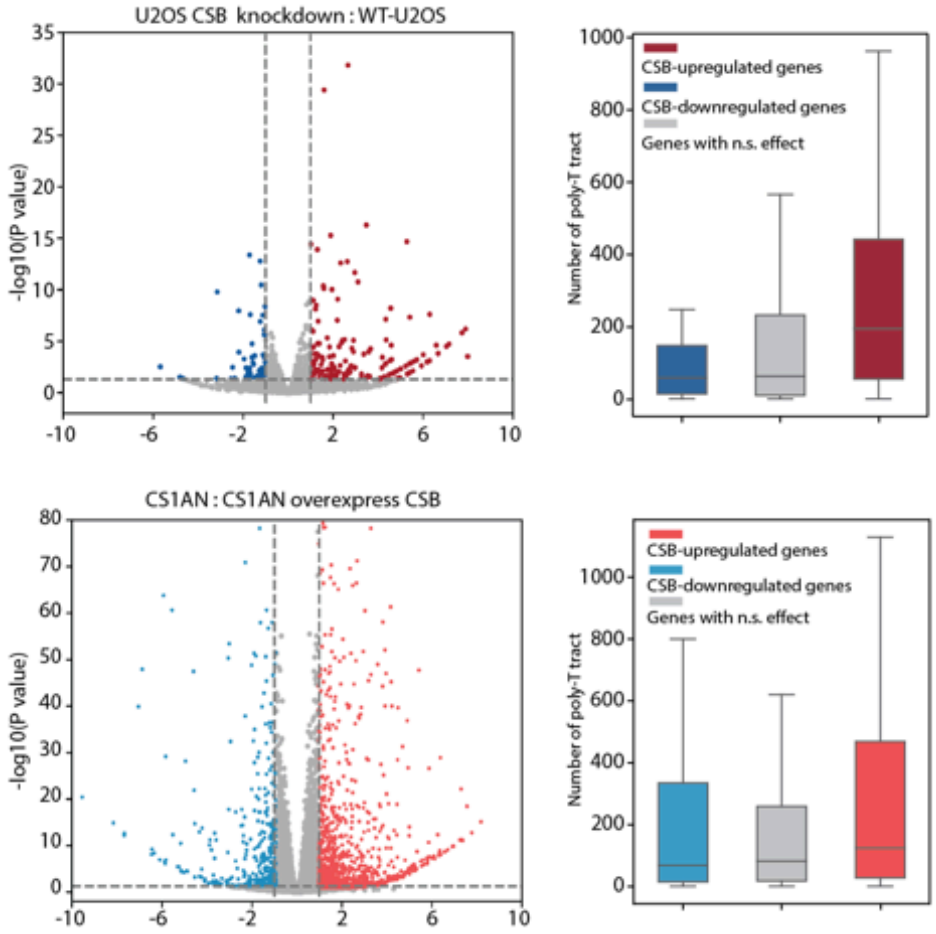

Figure 3.18: **Differential gene expression analysis on CSB deficient vs. CSB expressing.** (Top) (Left) Volcano plot of differential gene expression analysis on CSB knockdown vs normal condition in U2OS cell. (Right) The number of poly T tract in each gene in three sets, CSB expression downregulated genes (dark blue), CSB expression insensitive genes (grey) and CSB expression upregulated genes (Dark red) (Bottom) Volcano plot of differential gene expression analysis on CS1AN (Cockayne Syndrome patient cell with CSB deficient) vs CS1AN cell overexpess CSB. (Right) The number of poly T tract in each gene in three sets, CSB expression downregulated genes (light blue), CSB expression insensitive genes (grey) and CSB expression upregulated genes (light red).

.

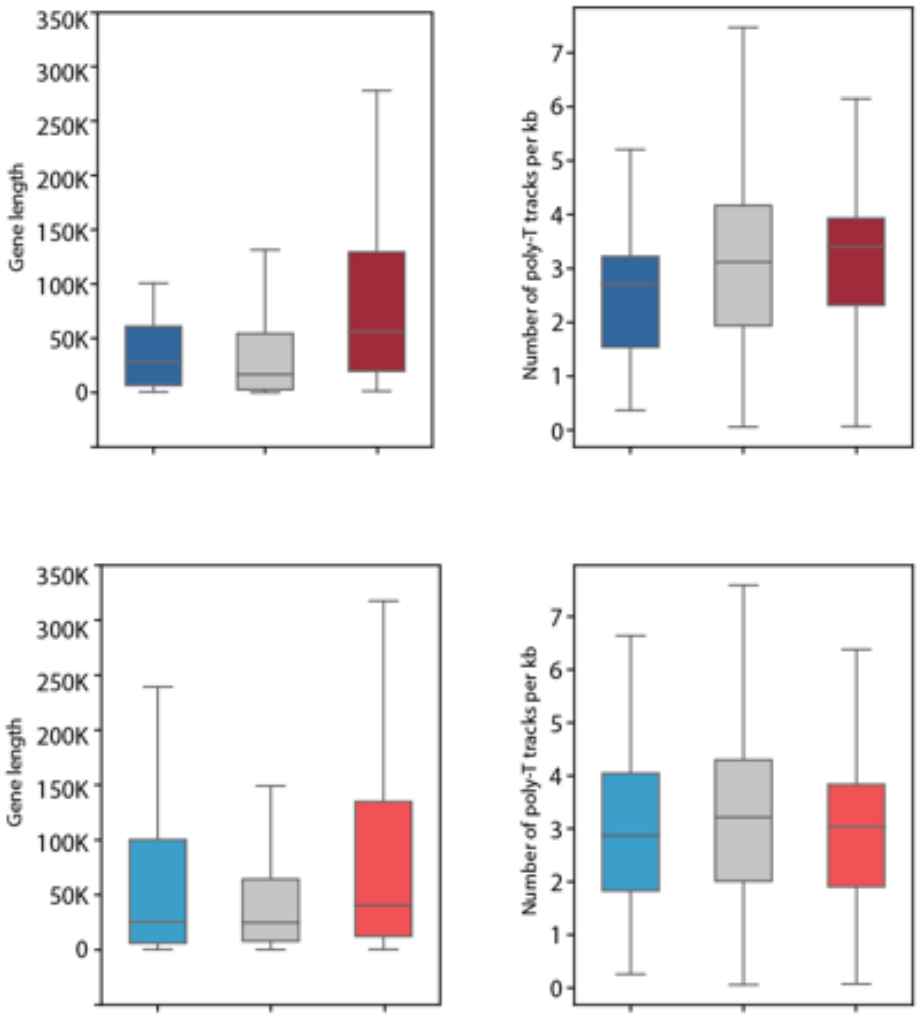

Figure 3.19: **Enrichment of long genes in CSB expression upregulated gene set.** (Top) (Left) Box plot of gene length in three gene sets, CSB expression downregulated genes (Dark blue), CSB expression insensitive genes (Grey) and CSB expression upregulated genes (Dark red). (Right) The number of poly T tracts per kilo base in each gene in three sets, CSB expression downregulated genes (Dark blue), CSB expression insensitive genes (grey) and CSB expression upregulated genes (Dark red) (Bottom) (Left) Box plot of gene length in three gene sets, CSB expression downregulated genes (Light blue), CSB expression insensitive genes (Grey) and CSB expression upregulated genes ((Light red). (Right) The number of poly T tracts per kilo base in each gene in three sets, CSB expression downregulated genes ((Light blue), CSB expression insensitive genes (grey) and CSB expression upregulated genes (Light red)

.

Chapter 4

Conclusions and future perspectives

4.1 Conclusion

Rencent studies have discovered the important roles of R-loops in gene expression regulation during development and diseases. To reveal the underlying regulatory nechanisms, the first step is to locate the potision of R-loops in vivo. DRIP-seq, relying on high affinity of the S9.6 monoclonal antibody for DNA:RNA hybrids, was first used for genome-wide profiling of R-loops in humans. While DRIP-seq provided rich information on R-loops, this technique suffers from limited resolution due to the fact that restriction enzymes are used to achieve DNA fragmentation. Also, for a long time, because of the absence of another antibody-based method for R-loop immunoprecipitation, validation of DRIP-seq results is difficult. In this thesis, we introduced the development of a high resolution R-loop mapping method, R-ChIP, which adopts the inactive RNase H1 instead of S9.6 to capture R-loops.

The following conclusions for R-ChIP can be drawn: A V5 tagged inactive RNase H1 was constructed, which is proven to achieve high affinity and specificity in capturing R-loops. The use of overexpressing mutant RNase H1 enables in vivo native R-loop profiling and probing more accurate R-loops dynamics, even for the detection of unstable and transient R-loops formed during transcription. Crosslinking and sonication replaced the restriction enzyme digestion to fragmentize DNA, which greatly promotes the resolution of R-loop position identification. Moreover, the final library was constructed in a strand-specific manner to provide the direction of the RNA moiety within R-loops, which further solidified our understanding of how R-loop is formed during transcription.

Similar to DRIP-seq, the majority of R-loops identified by R-ChIP are located at promoter regions of actively transcribed genes. Those R-loops' regulation are highly intertwined with transcriptional pause and release at promoter-proximal regions. We found that the higher the original R-loop level is, the less nascent RNA could be further induced at promoters. However, we also found that R-loop resolution is not a prerequisite for transcriptional pause release, suggesting that R-loop resolution and transcriptional pause release are independently regulated processes in the cell. To investigate how R-loop functions during transcription, we perfromed R-ChIP by perturbing the expression of a transcription elongation factor, CSB. Interestingly, we found that more R-loops are induced at gene body and a new feature, poly T tract on non-template strand, emerged when CSB is depleted. By a careful analysis on the transcription patterns and in vitro transcription validation on the poly T R-loops, we found that poly T R-loops are caused by Pol II pausing while CSB is one of the key factors that prevent Pol II pausing during transcription. These findings have important implications in understanding the mechanism of Cockayne Syndrom.

By applying inactive RNase H1, we could not only extract the R-loop postion information on the genome, but also identify the R-loop associated proteome. Beyond the R-ChIP, we also identified the R-loop proteome in human cells through the mass spectormetry. Taken advantage of our unbinding and uncut mutant RNase H1 (WKKDD) and unbinding only mutant RNase H1 (D210N), a high confident list of proteins have been identified which may get involved and participated in R-loop related biological processes. The R-loop proteome (Table 4.1) reveals genes involved in fundamental biological pathways, including rRNA processing, mRNA splicing, termination of transcription and mRNA 3' end processing. A series of disease related proteins have also been enriched with high priority in our list, such as DDX41 which suppresses R-loop levels and inflammatory signaling in Hematopoietic stem and progenitor cells production [Weinreb et al., 2021].

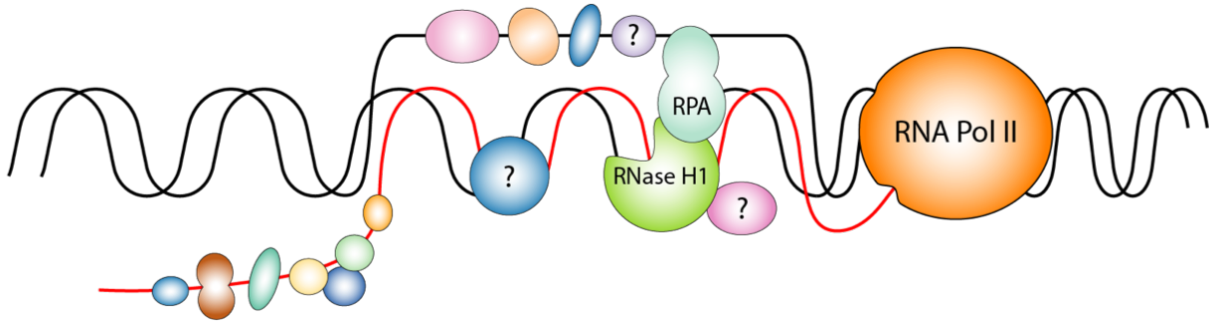

Figure 4.1: Schematic of R-loop interactome identification by RNase H1.

4.2 Future perspectives

PolyT induced R-loops may represent an important subset of co-transcriptionally induced R-loops. There is still a lot to be explored to explain the correlation between R-loops and pathogenic consequences. Fine mapping of native and induced R-loops is the key to unlock

how R-loops are regulated and how R-loops interfere with other biological processes. Also, by instituting different mutants of RNase H1, we may manipulate the level of R-loops in cells, making it possible for us to test the cause and effect of R-loops in genomic instability, development and diseases. In addition, the identified R-loop proteome pave the way for advancing R-loop biology in diverse developmental and disease processes.

Table 4.1: **R-loop interactome identified by RNase H1 mutants followed by mass spectrometry.** Fold change represents the protein intensity ratio between RNase H1 mutant D210N signal over WKKDD signal.

Gene Name	Description	Fold Change
DDX41	Probable ATP-dependent RNA helicase DDX41 OS=Homo sapiens OX=9606 GN=DDX41 PE=1 SV=2	37.97468354
PICALM	Phosphatidylinositol-binding clathrin assembly protein (Fragment) OS=Homo sapiens OX=9606 GN=PICALM PE=1 SV=1	17.41746497
RNF20	E3 ubiquitin-protein ligase BRE1A OS=Homo sapiens OX=9606 GN=RNF20 PE=1 SV=2	17.31016296
PGP	Glycerol-3-phosphate phosphatase OS=Homo sapiens OX=9606 GN=PGP PE=1 SV=1	16.26280696
RPA1	Replication protein A 70 kDa DNA-binding subunit OS=Homo sapiens OX=9606 GN=RPA1 PE=1 SV=2	12.50317109
EIF2B5	Translation initiation factor eIF-2B subunit epsilon OS=Homo sapiens OX=9606 GN=EIF2B5 PE=1 SV=1	12.08463608
SPATA5	ATPase family protein 2 homolog OS=Homo sapiens OX=9606 GN=SPATA5 PE=1 SV=3	11.70093458

Continued on next page

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
YTHDC2	3'-5' RNA helicase YTHDC2 OS=Homo sapiens OX=9606 GN=YTHDC2 PE=1 SV=2	10.68181818
MAPK1IP1L	MAPK-interacting and spindle-stabilizing protein-like OS=Homo sapiens OX=9606 GN=MAPK1IP1L PE=1 SV=4	10.0974026
CLUH	Clustered mitochondria protein homolog OS=Homo sapiens OX=9606 GN=CLUH PE=1 SV=1	8.75862069
RRM1	Ribonucleoside-diphosphate reductase large subunit OS=Homo sapiens OX=9606 GN=RRM1 PE=1 SV=1	8.537922606
EZH2	Histone-lysine N-methyltransferase EZH2 OS=Homo sapiens OX=9606 GN=EZH2 PE=1 SV=2	8.456096382
RTCA	RNA 3'-terminal phosphate cyclase OS=Homo sapiens OX=9606 GN=RTCA PE=1 SV=1	8.342051275
CANX	Calnexin OS=Homo sapiens OX=9606 GN=CANX PE=1 SV=2	8.214285714

Continued on next page

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
CKAP5	Cytoskeleton-associated protein 5 OS=Homo sapiens OX=9606 GN=CKAP5 PE=1 SV=3	8.063492063
SREK1IP1	Protein SREK1IP1 OS=Homo sapiens OX=9606 GN=SREK1IP1 PE=1 SV=1	7.431239679
PURB	Transcriptional activator protein Pur-beta OS=Homo sapiens OX=9606 GN=PURB PE=1 SV=3	7.239699882
YWHAB	14-3-3 protein beta/alpha OS=Homo sapiens OX=9606 GN=YWHAB PE=1 SV=3	7.232704403
OTUB1	Ubiquitin thioesterase OTUB1 OS=Homo sapiens OX=9606 GN=OTUB1 PE=1 SV=1	7.063636364
ADRM1	Proteasomal ubiquitin receptor ADRM1 OS=Homo sapiens OX=9606 GN=ADRM1 PE=1 SV=2	6.806043682
ELAC2	Zinc phosphodiesterase ELAC protein 2 (Fragment) OS=Homo sapiens OX=9606 GN=ELAC2 PE=1 SV=2	6.663769376
COMT	Catechol O-methyltransferase OS=Homo sapiens OX=9606 GN=COMT PE=1 SV=2	6.253776435

Continued on next page

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
SLC4A1AP	Kanadaptin OS=Homo sapiens OX=9606 GN=SLC4A1AP PE=1 SV=1	6.231306082
ZNF318	Zinc finger protein 318 OS=Homo sapiens OX=9606 GN=ZNF318 PE=1 SV=2	5.856079404
AURKA	Aurora kinase A OS=Homo sapiens OX=9606 GN=AURKA PE=1 SV=2	5.754385965
HIST4H4	Histone H4 OS=Homo sapiens OX=9606 GN=HIST1H4A PE=1 SV=2	5.686015831
NOLC1	Nucleolar and coiled-body phosphoprotein 1 OS=Homo sapiens OX=9606 GN=NOLC1 PE=1 SV=2	5.282258065
MRE11	Double-strand break repair protein OS=Homo sapiens OX=9606 GN=MRE11 PE=1 SV=1	5.244094488
CDK11B	Cyclin-dependent kinase 11B OS=Homo sapiens OX=9606 GN=CDK11B PE=1 SV=1	5.203703704
GRSF1	G-rich sequence factor 1 OS=Homo sapiens OX=9606 GN=GRSF1 PE=1 SV=3	5.105386417

Continued on next page

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
RPSA	40S ribosomal protein SA (Fragment) OS=Homo sapiens OX=9606 GN=RPSA PE=1 SV=8	5.101214575
RANGAP1	Ran GTPase-activating protein 1 OS=Homo sapiens OX=9606 GN=RANGAP1 PE=1 SV=1	4.924528302
PSMF1	Proteasome inhibitor PI31 subunit OS=Homo sapiens OX=9606 GN=PSMF1 PE=1 SV=2	4.901655347
NKAP	NF-kappa-B-activating protein (Fragment) OS=Homo sapiens OX=9606 GN=NKAP PE=1 SV=1	4.841059603
CDC16	Cell division cycle protein 16 homolog OS=Homo sapiens OX=9606 GN=CDC16 PE=1 SV=1	4.657894737
RFC3	Replication factor C subunit 3 OS=Homo sapiens OX=9606 GN=RFC3 PE=1 SV=2	4.647368421
NFX1	Transcriptional repressor NF-X1 OS=Homo sapiens OX=9606 GN=NFX1 PE=1 SV=2	4.645669291
Continued on next page		

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
HHEX	Hematopoietically-expressed homeobox protein HHEX OS=Homo sapiens OX=9606 GN=HHEX PE=1 SV=1	4.588607595
HSP90AA1	Heat shock protein HSP 90-alpha OS=Homo sapiens OX=9606 GN=HSP90AA1 PE=1 SV=5	4.528169014
KIF2C	Kinesin-like protein KIF2C OS=Homo sapiens OX=9606 GN=KIF2C PE=1 SV=2	4.523809524
H2AFZ	Histone H2A.Z OS=Homo sapiens OX=9606 GN=H2AFZ PE=1 SV=2	4.353448276
ZNF326	DBIRD complex subunit ZNF326 OS=Homo sapiens OX=9606 GN=ZNF326 PE=1 SV=1	4.189636163
ELOC	Elongin-C (Fragment) OS=Homo sapiens OX=9606 GN=ELOC PE=1 SV=1	4.147894646
LBR	Delta(14)-sterol reductase LBR OS=Homo sapiens OX=9606 GN=LBR PE=1 SV=2	4.082872928
BANF1	Barrier-to-autointegration factor OS=Homo sapiens OX=9606 GN=BANF1 PE=1 SV=1	4.049382716

Continued on next page

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
PFAS	Phosphoribosylformylglycinamidine synthase OS=Homo sapiens OX=9606 GN=PFAS PE=1 SV=4	4.034285714
ASPH	Aspartyl/asparaginyl beta-hydroxylase OS=Homo sapiens OX=9606 GN=ASPH PE=1 SV=3	3.9765625
SRP72	Signal recognition particle subunit SRP72 OS=Homo sapiens OX=9606 GN=SRP72 PE=1 SV=3	3.973063973
ORC1	Origin recognition complex subunit 1 OS=Homo sapiens OX=9606 GN=ORC1 PE=1 SV=2	3.947867299
HLTF	Helicase-like transcription factor OS=Homo sapiens OX=9606 GN=HLTF PE=1 SV=1	3.940594059
DNM2	Dynamin-2 OS=Homo sapiens OX=9606 GN=DNM2 PE=1 SV=2	3.908450704
CKAP2	Cytoskeleton-associated protein 2 OS=Homo sapiens OX=9606 GN=CKAP2 PE=1 SV=1	3.856115108
		Continued on next page

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
PMPCA	Mitochondrial-processing peptidase subunit alpha OS=Homo sapiens OX=9606 GN=PMPCA PE=1 SV=2	3.829832402
TARDBP	TAR DNA-binding protein 43 OS=Homo sapiens OX=9606 GN=TARDBP PE=1 SV=1	3.779005525
NFRKB	Nuclear factor related to kappa-B-binding protein OS=Homo sapiens OX=9606 GN=NFRKB PE=1 SV=2	3.740415186
CHD7	Chromodomain-helicase-DNA-binding protein 7 OS=Homo sapiens OX=9606 GN=CHD7 PE=1 SV=3	3.727969349
ASCC2	Activating signal cointegrator 1 complex subunit 2 OS=Homo sapiens OX=9606 GN=ASCC2 PE=1 SV=3	3.705035971
CYFIP2	Cytoplasmic FMR1-interacting protein 2 OS=Homo sapiens OX=9606 GN=CYFIP2 PE=1 SV=1	3.67826087
RPS6KB2	Ribosomal protein S6 kinase beta-2 OS=Homo sapiens OX=9606 GN=RPS6KB2 PE=1 SV=2	3.67816092

Continued on next page

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
APOBEC3C	DNA dC->dU-editing enzyme APOBEC-3C OS=Homo sapiens OX=9606 GN=APOBEC3C PE=1 SV=2	3.624338624
FECH	Ferrochelatase mitochondrial OS=Homo sapiens OX=9606 GN=FECH PE=1 SV=2	3.60522758
GIGYF2	GRB10-interacting GYF protein 2 OS=Homo sapiens OX=9606 GN=GIGYF2 PE=1 SV=1	3.571428571
PAK1IP1	p21-activated protein kinase-interacting protein 1 OS=Homo sapiens OX=9606 GN=PAK1IP1 PE=1 SV=2	3.513513514
ZBTB7A	Zinc finger and BTB domain-containing protein 7A OS=Homo sapiens OX=9606 GN=ZBTB7A PE=1 SV=1	3.378151261
BLM	Bloom syndrome protein OS=Homo sapiens OX=9606 GN=BLM PE=1 SV=1	3.362903226
EED	Polycomb protein EED OS=Homo sapiens OX=9606 GN=EED PE=1 SV=2	3.355191257

Continued on next page

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
POLR3A	DNA-directed RNA polymerase III subunit RPC1 OS=Homo sapiens OX=9606 GN=POLR3A PE=1 SV=2	3.326693227
DNAJC19	Mitochondrial import inner membrane translocase subunit TIM14 OS=Homo sapiens OX=9606 GN=DNAJC19 PE=1 SV=3	3.303482587
H3F3B	Histone H3 (Fragment) OS=Homo sapiens OX=9606 GN=H3F3B PE=1 SV=1	3.251908397
FARS2	Phenylalanine–tRNA ligase mitochondrial OS=Homo sapiens OX=9606 GN=FARS2 PE=1 SV=1	3.250773994
ZNF845	Zinc finger protein 845 OS=Homo sapiens OX=9606 GN=ZNF845 PE=1 SV=3	3.169354839
TRIM56	E3 ubiquitin-protein ligase TRIM56 OS=Homo sapiens OX=9606 GN=TRIM56 PE=1 SV=3	3.158415842
PLEKHA5	Pleckstrin homology domain-containing family A member 5 OS=Homo sapiens OX=9606 GN=PLEKHA5 PE=1 SV=1	3.140495868
Continued on next page		

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
IARS	Isoleucine–tRNA ligase cytoplasmic OS=Homo sapiens OX=9606 GN=IARS PE=1 SV=2	3.138157895
RHOG	Rho-related GTP-binding protein RhoG OS=Homo sapiens OX=9606 GN=RHOG PE=1 SV=1	3.075313808
PSMD4	26S proteasome non-ATPase regulatory subunit 4 OS=Homo sapiens OX=9606 GN=PSMD4 PE=1 SV=1	3.058252427
ASF1B	Histone chaperone ASF1B OS=Homo sapiens OX=9606 GN=ASF1B PE=1 SV=1	3.056603774
C19orf47	Uncharacterized protein C19orf47 OS=Homo sapiens OX=9606 GN=C19orf47 PE=1 SV=1	3.043209877
SPATS2L	SPATS2-like protein OS=Homo sapiens OX=9606 GN=SPATS2L PE=1 SV=2	3.033898305
DNMT1	DNA (cytosine-5)-methyltransferase 1 OS=Homo sapiens OX=9606 GN=DNMT1 PE=1 SV=2	3.031578947
OTUD4	OTU domain-containing protein 4 OS=Homo sapiens OX=9606 GN=OTUD4 PE=1 SV=4	3.005780347

Continued on next page

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
MEMO1	Protein MEMO1 OS=Homo sapiens OX=9606 GN=MEMO1 PE=1 SV=1	2.996019574
DNAJA2	DnaJ homolog subfamily A member 2 OS=Homo sapiens OX=9606 GN=DNAJA2 PE=1 SV=1	2.982142857
AFDN	Afadin (Fragment) OS=Homo sapiens OX=9606 GN=AFDN PE=1 SV=1	2.972972973
AP3D1	AP-3 complex subunit delta-1 OS=Homo sapiens OX=9606 GN=AP3D1 PE=1 SV=1	2.967244701
ZC3H8	Zinc finger CCCH domain-containing protein 8 OS=Homo sapiens OX=9606 GN=ZC3H8 PE=1 SV=2	2.96625222
CBX8	Chromobox protein homolog 8 OS=Homo sapiens OX=9606 GN=CBX8 PE=1 SV=3	2.939481268
PSMD11	26S proteasome non-ATPase regulatory subunit 11 (Fragment) OS=Homo sapiens OX=9606 GN=PSMD11 PE=1 SV=8	2.936026936
ABHD14B	Protein ABHD14B OS=Homo sapiens OX=9606 GN=ABHD14B PE=1 SV=1	2.926769532

Continued on next page

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
SNRNP48	U11/U12 small nuclear ribonucleoprotein 48 kDa protein OS=Homo sapiens OX=9606 GN=SNRNP48 PE=1 SV=2	2.904
YY1AP1	YY1-associated protein 1 OS=Homo sapiens OX=9606 GN=YY1AP1 PE=1 SV=2	2.898209741
RPS21	40S ribosomal protein S21 OS=Homo sapiens OX=9606 GN=RPS21 PE=1 SV=1	2.886363636
PLK1	Serine/threonine-protein kinase PLK1 OS=Homo sapiens OX=9606 GN=PLK1 PE=1 SV=1	2.880921895
LDHB	L-lactate dehydrogenase B chain OS=Homo sapiens OX=9606 GN=LDHB PE=1 SV=2	2.878504673
CLTC	Clathrin heavy chain 1 OS=Homo sapiens OX=9606 GN=CLTC PE=1 SV=5	2.816993464
SF3B3	Splicing factor 3B subunit 3 OS=Homo sapiens OX=9606 GN=SF3B3 PE=1 SV=4	2.814371257
HMGA1	High mobility group protein HMG-I/HMG-Y OS=Homo sapiens OX=9606 GN=HMGA1 PE=1 SV=3	2.810902896

Continued on next page

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
NUMA1	Nuclear mitotic apparatus protein 1 OS=Homo sapiens OX=9606 GN=NUMA1 PE=1 SV=1	2.805280528
ZW10	Centromere/kinetochore protein zw10 homolog OS=Homo sapiens OX=9606 GN=ZW10 PE=1 SV=3	2.804216867
GRB2	Growth factor receptor-bound protein 2 (Fragment) OS=Homo sapiens OX=9606 GN=GRB2 PE=1 SV=8	2.781021898
RBM4	RNA-binding protein 4 OS=Homo sapiens OX=9606 GN=RBM4 PE=1 SV=1	2.780254777
HIST1H2BK	Histone H2B type 1-K OS=Homo sapiens OX=9606 GN=HIST1H2BK PE=1 SV=3	2.777777778
AKAP9	A-kinase anchor protein 9 OS=Homo sapiens OX=9606 GN=AKAP9 PE=1 SV=4	2.769668388
RRP1	Ribosomal RNA processing protein 1 homolog A OS=Homo sapiens OX=9606 GN=RRP1 PE=1 SV=1	2.73556231
SUPT16H	FACT complex subunit SPT16 OS=Homo sapiens OX=9606 GN=SUPT16H PE=1 SV=1	2.725

Continued on next page

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
ELOB	Elongin-B OS=Homo sapiens OX=9606 GN=ELOB PE=1 SV=1	2.717948718
TOP2B	DNA topoisomerase 2 (Fragment) OS=Homo sapiens OX=9606 GN=TOP2B PE=1 SV=1	2.697686464
MAP3K20	Mitogen-activated protein kinase kinase kinase 20 OS=Homo sapiens OX=9606 GN=MAP3K20 PE=1 SV=3	2.695652174
KNSTRN	Small kinetochore-associated protein OS=Homo sapiens OX=9606 GN=KNSTRN PE=1 SV=2	2.678571429
MORC3	MORC family CW-type zinc finger protein 3 OS=Homo sapiens OX=9606 GN=MORC3 PE=1 SV=3	2.669376694
AIMP1	Aminoacyl tRNA synthase complex-interacting multifunctional protein 1 OS=Homo sapiens OX=9606 GN=AIMP1 PE=1 SV=2	2.661341853
RIOK2	Serine/threonine-protein kinase RIO2 OS=Homo sapiens OX=9606 GN=RIOK2 PE=1 SV=2	2.643678161

Continued on next page

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
OSTC	Oligosaccharyltransferase complex subunit OSTC OS=Homo sapiens OX=9606 GN=OSTC PE=1 SV=1	2.6435247
SMC4	Structural maintenance of chromosomes protein 4 OS=Homo sapiens OX=9606 GN=SMC4 PE=1 SV=2	2.622036262
BOP1	Ribosome biogenesis protein BOP1 OS=Homo sapiens OX=9606 GN=BOP1 PE=1 SV=2	2.616926503
KIF23	Kinesin-like protein OS=Homo sapiens OX=9606 GN=KIF23 PE=1 SV=1	2.608465608
CAD	CAD protein OS=Homo sapiens OX=9606 GN=CAD PE=1 SV=1	2.607223476
GTF2I	General transcription factor II-I OS=Homo sapiens OX=9606 GN=GTF2I PE=1 SV=2	2.590909091
NSA2	Ribosome biogenesis protein NSA2 homolog OS=Homo sapiens OX=9606 GN=NSA2 PE=1 SV=1	2.590759076

Continued on next page

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
GMPS	GMP synthase [glutamine-hydrolyzing] OS=Homo sapiens OX=9606 GN=GMPS PE=1 SV=1	2.581632653
PDIA6	Protein disulfide-isomerase A6 OS=Homo sapiens OX=9606 GN=PDIA6 PE=1 SV=1	2.568807339
TUFM	Elongation factor Tu mitochondrial OS=Homo sapiens OX=9606 GN=TUFM PE=1 SV=2	2.560283688
ATP1A1	Sodium/potassium-transporting ATPase subunit alpha-1 OS=Homo sapiens OX=9606 GN=ATP1A1 PE=1 SV=1	2.550200803
RBM28	RNA-binding protein 28 OS=Homo sapiens OX=9606 GN=RBM28 PE=1 SV=3	2.524038462
BAZ1B	Tyrosine-protein kinase BAZ1B OS=Homo sapiens OX=9606 GN=BAZ1B PE=1 SV=2	2.514851485
RRP1B	Ribosomal RNA processing protein 1 homolog B OS=Homo sapiens OX=9606 GN=RRP1B PE=1 SV=3	2.508403361
PPM1G	Protein phosphatase 1G OS=Homo sapiens OX=9606 GN=PPM1G PE=1 SV=1	2.507352941

Continued on next page

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
STK26	Serine/threonine-protein kinase 26 OS=Homo sapiens OX=9606 GN=STK26 PE=1 SV=1	2.502720348
NOP53	Ribosome biogenesis protein NOP53 OS=Homo sapiens OX=9606 GN=NOP53 PE=1 SV=2	2.493333333
RARS	Arginine–tRNA ligase cytoplasmic OS=Homo sapiens OX=9606 GN=RARS PE=1 SV=2	2.490566038
HNRNPUL2-BSCL2	HCG2044799 OS=Homo sapiens OX=9606 GN=HNRNPUL2-BSCL2 PE=4 SV=1	2.474074074
RSBN1	Lysine-specific demethylase 9 OS=Homo sapiens OX=9606 GN=RSBN1 PE=1 SV=1	2.472392638
DNAJA1	DnaJ homolog subfamily A member 1 OS=Homo sapiens OX=9606 GN=DNAJA1 PE=1 SV=2	2.46373057
SNRPC	U1 small nuclear ribonucleoprotein C OS=Homo sapiens OX=9606 GN=SNRPC PE=1 SV=1	2.461832061
MCM4	DNA helicase OS=Homo sapiens OX=9606 GN=MCM4 PE=1 SV=1	2.454918033
Continued on next page		

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
PPM1F	Protein phosphatase 1F OS=Homo sapiens OX=9606 GN=PPM1F PE=1 SV=1	2.428571429
BAG4	BAG family molecular chaperone regulator 4 OS=Homo sapiens OX=9606 GN=BAG4 PE=1 SV=1	2.423529412
VASP	Vasodilator-stimulated phosphoprotein OS=Homo sapiens OX=9606 GN=VASP PE=1 SV=3	2.414578588
AASDHPPT	L-aminoadipate-semialdehyde dehydrogenase-phosphopantetheinyl transferase OS=Homo sapiens OX=9606 GN=AASDHPPT PE=1 SV=2	2.408163265
DHX37	Probable ATP-dependent RNA helicase DHX37 OS=Homo sapiens OX=9606 GN=DHX37 PE=1 SV=1	2.407079646
LGALS1	Galectin-1 OS=Homo sapiens OX=9606 GN=LGALS1 PE=1 SV=2	2.403598972

Continued on next page

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
DDX50	ATP-dependent RNA helicase DDX50 OS=Homo sapiens OX=9606 GN=DDX50 PE=1 SV=1	2.403508772
GRWD1	Glutamate-rich WD repeat-containing protein 1 OS=Homo sapiens OX=9606 GN=GRWD1 PE=1 SV=1	2.381889764
MAP2K7	Dual specificity mitogen-activated protein kinase kinase 7 OS=Homo sapiens OX=9606 GN=MAP2K7 PE=1 SV=2	2.381868132
EDC4	Enhancer of mRNA-decapping protein 4 OS=Homo sapiens OX=9606 GN=EDC4 PE=1 SV=1	2.379032258
ERI1	3'-5' exoribonuclease 1 OS=Homo sapiens OX=9606 GN=ERI1 PE=1 SV=3	2.362962963
VARS	Valine–tRNA ligase OS=Homo sapiens OX=9606 GN=VARS PE=1 SV=4	2.359307359
ANKZF1	Ankyrin repeat and zinc finger domain-containing protein 1 OS=Homo sapiens OX=9606 GN=ANKZF1 PE=1 SV=1	2.355555556

Continued on next page

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
ESF1	ESF1 homolog OS=Homo sapiens OX=9606 GN=ESF1 PE=1 SV=1	2.353612167
LENG8	Leukocyte receptor cluster member 8 OS=Homo sapiens OX=9606 GN=LENG8 PE=1 SV=3	2.352941176
PRRC2B	Protein PRRC2B OS=Homo sapiens OX=9606 GN=PRRC2B PE=1 SV=2	2.350282486
MEPCE	7SK snRNA methylphosphate capping enzyme OS=Homo sapiens OX=9606 GN=MEPCE PE=1 SV=1	2.337398374
MAP1A	Microtubule-associated protein 1A OS=Homo sapiens OX=9606 GN=MAP1A PE=1 SV=6	2.327698309
SIRT1	NAD-dependent protein deacetylase sirtuin-1 OS=Homo sapiens OX=9606 GN=SIRT1 PE=1 SV=2	2.321564549
LARS	Leucine–tRNA ligase cytoplasmic OS=Homo sapiens OX=9606 GN=LARS PE=1 SV=2	2.30273752
MARS	Methionine–tRNA ligase cytoplasmic OS=Homo sapiens OX=9606 GN=MARS PE=1 SV=2	2.294277929

Continued on next page

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
ZGPAT	Zinc finger CCCH-type with G patch domain-containing protein OS=Homo sapiens OX=9606 GN=ZGPAT PE=1 SV=3	2.292857143
EIF4A1	Eukaryotic initiation factor 4A-I OS=Homo sapiens OX=9606 GN=EIF4A1 PE=1 SV=1	2.291457286
SRP68	Signal recognition particle subunit SRP68 OS=Homo sapiens OX=9606 GN=SRP68 PE=1 SV=2	2.286751361
EPRS	Bifunctional glutamate/proline–tRNA ligase OS=Homo sapiens OX=9606 GN=EPRS PE=1 SV=5	2.253829322
RPL39	60S ribosomal protein L39 OS=Homo sapiens OX=9606 GN=RPL39 PE=1 SV=2	2.25
FHOD1	FH1/FH2 domain-containing protein 1 OS=Homo sapiens OX=9606 GN=FHOD1 PE=1 SV=3	2.25
DDX10	Probable ATP-dependent RNA helicase DDX10 OS=Homo sapiens OX=9606 GN=DDX10 PE=1 SV=2	2.248466258

Continued on next page

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
ATP5PB	ATP synthase F(0) complex subunit B1 mitochondrial OS=Homo sapiens OX=9606 GN=ATP5PB PE=1 SV=1	2.241025641
MTHFD1	C-1-tetrahydrofolate synthase cytoplasmic (Fragment) OS=Homo sapiens OX=9606 GN=MTHFD1 PE=1 SV=1	2.237623762
ACOT7	Cytosolic acyl coenzyme A thioester hydrolase OS=Homo sapiens OX=9606 GN=ACOT7 PE=1 SV=3	2.23709369
SF3B1	Splicing factor 3B subunit 1 OS=Homo sapiens OX=9606 GN=SF3B1 PE=1 SV=3	2.236619718
PTPN1	Tyrosine-protein phosphatase non-receptor type 1 OS=Homo sapiens OX=9606 GN=PTPN1 PE=1 SV=1	2.227722772
AKAP8L	A-kinase anchor protein 8-like OS=Homo sapiens OX=9606 GN=AKAP8L PE=1 SV=4	2.221854305
PPP1R8	Nuclear inhibitor of protein phosphatase 1 OS=Homo sapiens OX=9606 GN=PPP1R8 PE=1 SV=2	2.206730769

Continued on next page

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
AFG3L2	AFG3-like protein 2 OS=Homo sapiens OX=9606 GN=AFG3L2 PE=1 SV=2	2.206477733
TIMM44	Mitochondrial import inner membrane translocase subunit TIM44 (Fragment) OS=Homo sapiens OX=9606 GN=TIMM44 PE=1 SV=1	2.204585538
ACAT1	Acetyl-CoA acetyltransferase mitochondrial OS=Homo sapiens OX=9606 GN=ACAT1 PE=1 SV=1	2.203791469
TUBA4A	Tubulin alpha-4A chain OS=Homo sapiens OX=9606 GN=TUBA4A PE=1 SV=1	2.2
DARS	Aspartate–tRNA ligase cytoplasmic OS=Homo sapiens OX=9606 GN=DARS PE=1 SV=2	2.192448234
WDR36	WD repeat-containing protein 36 OS=Homo sapiens OX=9606 GN=WDR36 PE=1 SV=1	2.183673469
CDK1	Cyclin-dependent kinase 1 OS=Homo sapiens OX=9606 GN=CDK1 PE=1 SV=3	2.177570093
HBZ	Hemoglobin subunit zeta OS=Homo sapiens OX=9606 GN=HBZ PE=1 SV=2	2.173913043
Continued on next page		

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
KARS	Lysine–tRNA ligase OS=Homo sapiens OX=9606 GN=KARS PE=1 SV=3	2.171511628
MSL1	Male-specific lethal 1 homolog OS=Homo sapiens OX=9606 GN=MSL1 PE=1 SV=1	2.162601626
CCNB1	G2/mitotic-specific cyclin-B1 OS=Homo sapiens OX=9606 GN=CCNB1 PE=1 SV=1	2.161328671
FAHD2A	Fumarylacetoacetate hydrolase domain-containing protein 2A (Fragment) OS=Homo sapiens OX=9606 GN=FAHD2A PE=1 SV=1	2.160714286
EFHD2	EF-hand domain-containing protein D2 (Fragment) OS=Homo sapiens OX=9606 GN=EFHD2 PE=1 SV=1	2.158119658
SRP14	Signal recognition particle 14 kDa protein OS=Homo sapiens OX=9606 GN=SRP14 PE=1 SV=2	2.155309033
PPP1R10	Serine/threonine-protein phosphatase 1 regulatory subunit 10 OS=Homo sapiens OX=9606 GN=PPP1R10 PE=1 SV=1	2.15503876

Continued on next page

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
BLVRA	Biliverdin reductase A OS=Homo sapiens OX=9606 GN=BLVRA PE=1 SV=2	2.149606299
ARHGEF2	Rho guanine nucleotide exchange factor 2 OS=Homo sapiens OX=9606 GN=ARHGEF2 PE=1 SV=4	2.146341463
TDRD3	Tudor domain-containing protein 3 OS=Homo sapiens OX=9606 GN=TDRD3 PE=1 SV=1	2.145969737
NPEPPS	Aminopeptidase OS=Homo sapiens OX=9606 GN=NPEPPS PE=1 SV=1	2.139744787
LUC7L	Putative RNA-binding protein Luc7-like 1 OS=Homo sapiens OX=9606 GN=LUC7L PE=1 SV=1	2.136363636
MACROD1	ADP-ribose glycohydrolase MACROD1 OS=Homo sapiens OX=9606 GN=MACROD1 PE=1 SV=2	2.134860051
SF3B4	Splicing factor 3B subunit 4 OS=Homo sapiens OX=9606 GN=SF3B4 PE=1 SV=1	2.132963989

Continued on next page

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
GPKOW	G-patch domain and KOW motifs-containing protein OS=Homo sapiens OX=9606 GN=GPKOW PE=1 SV=2	2.131578947
TMCO1	Calcium load-activated calcium channel (Fragment) OS=Homo sapiens OX=9606 GN=TMCO1 PE=1 SV=1	2.129770992
HOXB4	Homeobox protein Hox-B4 OS=Homo sapiens OX=9606 GN=HOXB4 PE=1 SV=2	2.129277567
ZC3H7B	Zinc finger CCCH domain-containing protein 7B OS=Homo sapiens OX=9606 GN=ZC3H7B PE=1 SV=2	2.12173913
ATAD3A	ATPase family AAA domain-containing protein 3A OS=Homo sapiens OX=9606 GN=ATAD3A PE=1 SV=2	2.120567376
PSMD5	26S proteasome non-ATPase regulatory subunit 5 OS=Homo sapiens OX=9606 GN=PSMD5 PE=1 SV=3	2.118483412

Continued on next page

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
PRKAR2B	cAMP-dependent protein kinase type II-beta regulatory subunit OS=Homo sapiens OX=9606 GN=PRKAR2B PE=1 SV=3	2.116504854
CEBPZ	CCAAT/enhancer-binding protein zeta OS=Homo sapiens OX=9606 GN=CEBPZ PE=1 SV=3	2.114537445
CMBL	Carboxymethylenebutenolidase homolog OS=Homo sapiens OX=9606 GN=CMBL PE=1 SV=1	2.112449799
NOM1	Nucleolar MIF4G domain-containing protein 1 OS=Homo sapiens OX=9606 GN=NOM1 PE=1 SV=1	2.10421286
RUVBL2	RuvB-like 2 OS=Homo sapiens OX=9606 GN=RUVBL2 PE=1 SV=3	2.097649186
HEMGN	Hemogen OS=Homo sapiens OX=9606 GN=HEMGN PE=1 SV=1	2.094488189
RANBP1	Ran-specific GTPase-activating protein OS=Homo sapiens OX=9606 GN=RANBP1 PE=1 SV=1	2.089761571

Continued on next page

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
DYNLL2	Dynein light chain 2 cytoplasmic OS=Homo sapiens OX=9606 GN=DYNLL2 PE=1 SV=1	2.089005236
HACD3	Very-long-chain (3R)-3-hydroxyacyl-CoA dehydratase 3 OS=Homo sapiens OX=9606 GN=HACD3 PE=1 SV=2	2.088757396
DECR1	2 4-dienoyl-CoA reductase mitochondrial OS=Homo sapiens OX=9606 GN=DECR1 PE=1 SV=1	2.083682008
FARSB	Phenylalanine–tRNA ligase beta subunit OS=Homo sapiens OX=9606 GN=FARSB PE=1 SV=3	2.082872928
FTSJ3	pre-rRNA 2'-O-ribose RNA methyltrans-ferase FTSJ3 OS=Homo sapiens OX=9606 GN=FTSJ3 PE=1 SV=2	2.080536913
PPAT	Amidophosphoribosyltransferase OS=Homo sapiens OX=9606 GN=PPAT PE=1 SV=1	2.078431373
PPP1R9B	Neurabin-2 OS=Homo sapiens OX=9606 GN=PPP1R9B PE=1 SV=3	2.077244259
Continued on next page		

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
DDX56	Probable ATP-dependent RNA helicase DDX56 OS=Homo sapiens OX=9606 GN=DDX56 PE=1 SV=1	2.07293666
TMPO	Lamina-associated polypeptide 2 isoforms beta/gamma OS=Homo sapiens OX=9606 GN=TMPO PE=1 SV=2	2.069277108
RO60	60 kDa SS-A/Ro ribonucleoprotein (Fragment) OS=Homo sapiens OX=9606 GN=RO60 PE=1 SV=8	2.068259386
POLR3D	DNA-directed RNA polymerase III sub-unit RPC4 OS=Homo sapiens OX=9606 GN=POLR3D PE=1 SV=2	2.066666667
EEF2K	Eukaryotic elongation factor 2 kinase OS=Homo sapiens OX=9606 GN=EEF2K PE=1 SV=2	2.060810811
DLD	Dihydrolipoyl dehydrogenase mitochondrial OS=Homo sapiens OX=9606 GN=DLD PE=1 SV=2	2.056

Continued on next page

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
SAP18	Histone deacetylase complex subunit SAP18 OS=Homo sapiens OX=9606 GN=SAP18 PE=1 SV=1	2.052845528
CPSF3	Cleavage and polyadenylation specificity factor subunit 3 OS=Homo sapiens OX=9606 GN=CPSF3 PE=1 SV=1	2.052689352
RPS12	40S ribosomal protein S12 OS=Homo sapiens OX=9606 GN=RPS12 PE=1 SV=3	2.051597052
MT-CO2	Cytochrome c oxidase subunit 2 OS=Homo sapiens OX=9606 GN=MT-CO2 PE=1 SV=1	2.051212938
EEF1AKNMT	eEF1A lysine and N-terminal methyltransferase OS=Homo sapiens OX=9606 GN=EEF1AKNMT PE=1 SV=1	2.050980392
YWHAQ	14-3-3 protein theta OS=Homo sapiens OX=9606 GN=YWHAQ PE=1 SV=1	2.049947971
POP7	Ribonuclease P protein subunit p20 (Fragment) OS=Homo sapiens OX=9606 GN=POP7 PE=1 SV=1	2.045728039

Continued on next page

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
CSDE1	Cold shock domain-containing protein E1 OS=Homo sapiens OX=9606 GN=CSDE1 PE=1 SV=2	2.043422733
STIP1	Stress-induced-phosphoprotein 1 OS=Homo sapiens OX=9606 GN=STIP1 PE=1 SV=1	2.039155135
PRC1	Protein regulator of cytokinesis 1 OS=Homo sapiens OX=9606 GN=PRC1 PE=1 SV=2	2.033898305
POP1	Ribonucleases P/MRP protein subunit POP1 OS=Homo sapiens OX=9606 GN=POP1 PE=1 SV=2	2.02764977
SPATS2	Spermatogenesis-associated serine-rich protein 2 OS=Homo sapiens OX=9606 GN=SPATS2 PE=1 SV=1	2.027027027
PYM1	Partner of Y14 and mago OS=Homo sapiens OX=9606 GN=PYM1 PE=1 SV=1	2.026899351
IDH2	Isocitrate dehydrogenase [NADP] mitochon-drial OS=Homo sapiens OX=9606 GN=IDH2 PE=1 SV=2	2.023148148
Continued on next page		

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
RBM39	RNA-binding protein 39 OS=Homo sapiens OX=9606 GN=RBM39 PE=1 SV=2	2.020833333
ATP5F1D	ATP synthase subunit delta mitochondrial OS=Homo sapiens OX=9606 GN=ATP5F1D PE=1 SV=2	2.020497804
SUGT1	Protein SGT1 homolog OS=Homo sapiens OX=9606 GN=SUGT1 PE=1 SV=3	2.019480519
OXSR1	Serine/threonine-protein kinase OSR1 OS=Homo sapiens OX=9606 GN=OXSR1 PE=1 SV=1	2.010943912
SMCHD1	Structural maintenance of chromosomes flexible hinge domain-containing protein 1 OS=Homo sapiens OX=9606 GN=SMCHD1 PE=1 SV=2	2.006734007
EIF3A	Eukaryotic translation initiation factor 3 subunit A OS=Homo sapiens OX=9606 GN=EIF3A PE=1 SV=1	2.005319149
RPLP0	60S acidic ribosomal protein P0 OS=Homo sapiens OX=9606 GN=RPLP0 PE=1 SV=1	2.003231018

Continued on next page

Table 4.1 R-loop interactome identified by RNase H1 mutants, continued

Gene name	Description	Fold Change
MCM5	DNA replication licensing factor MCM5 OS=Homo sapiens OX=9606 GN=MCM5 PE=1 SV=5	2

Bibliography

[Bacolla et al., 2015] Bacolla, A., Wang, G., and Vasquez, K. M. (2015). New perspectives on dna and rna triplexes as effectors of biological activity. *PLoS genetics*, 11(12):e1005696.

[Bell et al., 2018] Bell, J. C., Jukam, D., Teran, N. A., Risca, V. I., Smith, O. K., Johnson, W. L., Skotheim, J. M., Greenleaf, W. J., and Straight, A. F. (2018). Chromatin-associated rna sequencing (char-seq) maps genome-wide rna-to-dna contacts. *Elife*, 7:e27024.

[Belotserkovskii et al., 2018] Belotserkovskii, B. P., Tornaletti, S., D'Souza, A. D., and Hanawalt, P. C. (2018). R-loop generation during transcription: Formation, processing and cellular outcomes. *DNA repair*, 71:69–81.

[Berget et al., 1977] Berget, S. M., Moore, C., and Sharp, P. A. (1977). Spliced segments at the 5′ terminus of adenovirus 2 late mrna. *Proceedings of the National Academy of Sciences*, 74(8):3171–3175.

[Chambers et al., 2015] Chambers, V. S., Marsico, G., Boutell, J. M., Di Antonio, M., Smith, G. P., and Balasubramanian, S. (2015). High-throughput sequencing of dna g-quadruplex structures in the human genome. *Nature biotechnology*, 33(8):877–881.

[Chédin et al., 2021] Chédin, F., Hartono, S. R., Sanz, L. A., and Vanoosthuyse, V. (2021). Best practices for the visualization, mapping, and manipulation of r-loops. *The EMBO Journal*, 40(4):e106394.

[Chen et al., 2019] Chen, J.-Y., Zhang, X., Fu, X.-D., and Chen, L. (2019). R-chip for genome-wide mapping of r-loops by using catalytically inactive rnaseh1. *Nature protocols*, 14(5):1661–1685.

[Chen et al., 2018] Chen, L., Chen, J.-Y., Huang, Y.-J., Gu, Y., Qiu, J., Qian, H., Shao, C., Zhang, X., Hu, J., Li, H., He, S., Zhou, Y., Abdel-Wahab, O., Zhang, D.-E., and Fu, X.-D. (2018). The

augmented r-loop is a unifying mechanism for myelodysplastic syndromes induced by high-risk splicing factor mutations. *Molecular cell*, 69(3):412–425.

[Chen et al., 2017] Chen, L., Chen, J.-Y., Zhang, X., Gu, Y., Xiao, R., Shao, C., Tang, P., Qian, H., Luo, D., Li, H., Zhou, Y., Zhang, D.-E., and Fu, X.-D. (2017). R-chip using inactive rnase h reveals dynamic coupling of r-loops with transcriptional pausing at gene promoters. *Molecular cell*, 68(4):745–757.

[Chen et al., 2015a] Chen, P. B., Chen, H. V., Acharya, D., Rando, O. J., and Fazzio, T. G. (2015a). R loops regulate promoter-proximal chromatin architecture and cellular differentiation. *Nature structural & molecular biology*, 22(12):999–1007.

[Chen et al., 2015b] Chen, P. B., Chen, H. V., Acharya, D., Rando, O. J., and Fazzio, T. G. (2015b). R loops regulate promoter-proximal chromatin architecture and cellular differentiation. *Nature structural & molecular biology*, 22(12):999–1007.

[Chow et al., 1977] Chow, L. T., Gelinas, R. E., Broker, T. R., and Roberts, R. J. (1977). An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger rna. *Cell*, 12(1):1–8.

[Dumelie and Jaffrey, 2017] Dumelie, J. G. and Jaffrey, S. R. (2017). Defining the location of promoter-associated r-loops at near-nucleotide resolution using bisdrip-seq. *Elife*, 6:e28306.

[Duquette et al., 2004] Duquette, M. L., Handa, P., Vincent, J. A., Taylor, A. F., and Maizels, N. (2004). Intracellular transcription of g-rich dnas induces formation of g-loops, novel structures containing g4 dna. *Genes & development*, 18(13):1618–1629.

[Dvinge et al., 2016] Dvinge, H., Kim, E., Abdel-Wahab, O., and Bradley, R. K. (2016). Rna splicing factors as oncoproteins and tumour suppressors. *Nature Reviews Cancer*, 16(7):413–430.

[D'Alessandro et al., 2018] D'Alessandro, G., Whelan, D. R., Howard, S. M., Vitelli, V., Renaudin, X., Adamowicz, M., Iannelli, F., Jones-Weinert, C. W., Lee, M., and Matti, V. (2018). Brca2 controls dna: Rna hybrid level at dsbs by mediating rnase h2 recruitment. *Nature communications*, 9(1):1–17.

[Feng et al., 2012] Feng, J., Liu, T., Qin, B., Zhang, Y., and Liu, X. S. (2012). Identifying chip-seq enrichment using macs. *Nature protocols*, 7(9):1728–1740.

[Fong et al., 2009] Fong, N., Öhman, M., and Bentley, D. L. (2009). Fast ribozyme cleavage releases transcripts from rna polymerase ii and aborts co-transcriptional pre-mrna processing. *Nature structural & molecular biology*, 16(9):916–922.

[Gabel et al., 2015] Gabel, H. W., Kinde, B., Stroud, H., Gilbert, C. S., Harmin, D. A., Kastan, N. R., Hemberg, M., Ebert, D. H., and Greenberg, M. E. (2015). Disruption of dna-methylation-dependent long gene repression in rett syndrome. *Nature*, 522(7554):89–93.

[García-Benítez et al., 2017] García-Benítez, F., Gaillard, H., and Aguilera, A. (2017). Physical proximity of chromatin to nuclear pores prevents harmful r loop accumulation contributing to maintain genome stability. *Proceedings of the National Academy of Sciences*, 114(41):10942–10947.

[Ginno et al., 2013] Ginno, P. A., Lim, Y. W., Lott, P. L., Korf, I., and Chédin, F. (2013). Gc skew at the 5′ and 3′ ends of human genes links r-loop formation to epigenetic regulation and transcription termination. *Genome research*, 23(10):1590–1600.

[Ginno et al., 2012] Ginno, P. A., Lott, P. L., Christensen, H. C., Korf, I., and Chédin, F. (2012). R-loop formation is a distinctive characteristic of unmethylated human cpg island promoters. *Molecular cell*, 45(6):814–825.

[Halász et al., 2017] Halász, L., Karányi, Z., Boros-Oláh, B., Kuik-Rózsa, T., Sipos, É., Nagy, É., Mosolygó-L, Á., Mázló, A., Rajnavölgyi, É., and Halmos, G. (2017). Rna-dna hybrid (r-loop) immunoprecipitation mapping: an analytical workflow to evaluate inherent biases. *Genome research*, 27(6):1063–1073.

[Hanawalt and Spivak, 2008] Hanawalt, P. C. and Spivak, G. (2008). Transcription-coupled dna repair: two decades of progress and surprises. *Nature reviews Molecular cell biology*, 9(12):958–970.

[Harinarayanan and Gowrishankar, 2003] Harinarayanan, R. and Gowrishankar, J. (2003). Host factor titration by chromosomal r-loops as a mechanism for runaway plasmid replication in transcription termination-defective mutants of escherichia coli. *Journal of molecular biology*, 332(1):31–46.

[Hartono et al., 2018] Hartono, S. R., Malapert, A., Legros, P., Bernard, P., Chédin, F., and Vanoosthuyse, V. (2018). The affinity of the s9. 6 antibody for double-stranded rnas impacts the accurate mapping of r-loops in fission yeast. *Journal of molecular biology*, 430(3):272–284.

[Hawkes et al., 2016] Hawkes, E. J., Hennelly, S. P., Novikova, I. V., Irwin, J. A., Dean, C., and Sanbonmatsu, K. Y. (2016). Coolair antisense rnas form evolutionarily conserved elaborate secondary structures. *Cell reports*, 16(12):3087–3096.

[Hong et al., 1995] Hong, X., Cadwell, G., and Kogoma, T. (1995). Escherichia coli recg and reca proteins in r-loop formation. *The EMBO Journal*, 14(10):2385–2392.

[Jang et al., 2020] Jang, Y., Elsayed, Z., Eki, R., He, S., Du, K.-P., Abbas, T., and Kai, M. (2020). Intrinsically disordered protein rbm14 plays a role in generation of rna: Dna hybrids at double-strand break sites. *Proceedings of the National Academy of Sciences*, 117(10):5329–5338.

[Jonkers and Lis, 2015] Jonkers, I. and Lis, J. T. (2015). Getting up to speed with transcription elongation by rna polymerase ii. *Nature reviews Molecular cell biology*, 16(3):167–177.

[Karikkineth et al., 2017] Karikkineth, A. C., Scheibye-Knudsen, M., Fivenson, E., Croteau, D. L., and Bohr, V. A. (2017). Cockayne syndrome: clinical features, model systems and pathways. *Ageing research reviews*, 33:3–17.

[Kim et al., 2015] Kim, E., Ilagan, J. O., Liang, Y., Daubner, G. M., Lee, S. C.-W., Ramakrishnan, A., Li, Y., Chung, Y. R., Micol, J.-B., Murphy, M. E., and Abdel-Wahab, O. (2015). Srsf2 mutations contribute to myelodysplasia by mutant-specific effects on exon recognition. *Cancer cell*, 27(5):617–630.

[Kim et al., 1999] Kim, H.-D., Choe, J., and Seo, Y.-S. (1999). The sen1+ gene of schizosaccharomyces pombe, a homologue of budding yeast sen1, encodes an rna and dna helicase. *Biochemistry*, 38(44):14697–14710.

[Kim and Jinks-Robertson, 2012] Kim, N. and Jinks-Robertson, S. (2012). Transcription as a source of genome instability. *Nature Reviews Genetics*, 13(3):204–214.

[Komeno et al., 2015] Komeno, Y., Huang, Y.-J., Qiu, J., Lin, L., Xu, Y., Zhou, Y., Chen, L., Monterroza, D. D., Li, H., DeKelver, R. C., Yan, M., Fu, X.-D., and Zhang, D.-E. (2015). Srsf2 is essential for hematopoiesis, and its myelodysplastic syndrome-related mutations dysregulate alternative pre-mrna splicing. *Molecular and cellular biology*, 35(17):3071–3082.

[König et al., 2017] König, F., Schubert, T., and Längst, G. (2017). The monoclonal s9. 6 antibody exhibits highly variable binding affinities towards different r-loop sequences. *PLoS One*, 12(6):e0178875.

[Li and Manley, 2005] Li, X. and Manley, J. L. (2005). Inactivation of the sr protein splicing factor asf/sf2 results in genomic instability. *Cell*, 122(3):365–378.

[Li et al., 2017] Li, X., Zhou, B., Chen, L., Gou, L.-T., Li, H., and Fu, X.-D. (2017). Grid-seq reveals the global rna–chromatin interactome. *Nature biotechnology*, 35(10):940–950.

[Li et al., 2016] Li, Y., Syed, J., and Sugiyama, H. (2016). Rna-dna triplex formation by long noncoding rnas. *Cell chemical biology*, 23(11):1325–1333.

[Mondal et al., 2015] Mondal, T., Subhash, S., Vaid, R., Enroth, S., Uday, S., Reinius, B., Mitra,

S., Mohammed, A., James, A. R., and Hoberg, E. (2015). Meg3 long noncoding rna regulates the tgf-β pathway genes through formation of rna–dna triplex structures. *Nature communications*, 6(1):1–17.

[Nguyen et al., 2017] Nguyen, H. D., Yadav, T., Giri, S., Saez, B., Graubert, T. A., and Zou, L. (2017). Functions of replication protein a as a sensor of r loops and a regulator of rnaseh1. *Molecular cell*, 65(5):832–847.

[Niehrs and Luke, 2020] Niehrs, C. and Luke, B. (2020). Regulatory r-loops as facilitators of gene expression and genome stability. *Nature Reviews Molecular Cell Biology*, 21(3):167–178.

[Nowotny et al., 2007] Nowotny, M., Gaidamakov, S. A., Ghirlando, R., Cerritelli, S. M., Crouch, R. J., and Yang, W. (2007). Structure of human rnase h1 complexed with an rna/dna hybrid: insight into hiv reverse transcription. *Molecular cell*, 28(2):264–276.

[O'Leary et al., 2015] O'Leary, V. B., Ovsepian, S. V., Carrascosa, L. G., Buske, F. A., Radulovic, V., Niyazi, M., Moertl, S., Trau, M., Atkinson, M. J., and Anastasov, N. (2015). Particle, a triplex-forming long ncrna, regulates locus-specific methylation in response to low-dose irradiation. *Cell reports*, 11(3):474–485.

[Postepska-Igielska et al., 2015] Postepska-Igielska, A., Giwojna, A., Gasri-Plotnitsky, L., Schmitt, N., Dold, A., Ginsberg, D., and Grummt, I. (2015). Lncrna khps1 regulates expression of the proto-oncogene sphk1 via triplex-mediated changes in chromatin structure. *Molecular cell*, 60(4):626–636.

[Qiu et al., 2016] Qiu, J., Zhou, B., Thol, F., Zhou, Y., Chen, L., Shao, C., DeBoever, C., Hou, J., Li, H., Chaturvedi, A., Ganser, A., Bejar, R., Zhang, D.-E., Fu, X.-D., and Heuser, M. (2016). Distinct splicing signatures affect converged pathways in myelodysplastic syndrome patients carrying mutations in different splicing regulators. *Rna*, 22(10):1535–1549.

[Roy et al., 2008] Roy, D., Yu, K., and Lieber, M. R. (2008). Mechanism of r-loop formation at immunoglobulin class switch sequences. *Molecular and cellular biology*, 28(1):50–60.

[Santos-Pereira and Aguilera, 2015] Santos-Pereira, J. M. and Aguilera, A. (2015). R loops: new modulators of genome dynamics and function. *Nature Reviews Genetics*, 16(10):583–597.

[Sanz et al., 2016] Sanz, L. A., Hartono, S. R., Lim, Y. W., Steyaert, S., Rajpurkar, A., Ginno, P. A., Xu, X., and Chédin, F. (2016). Prevalent, dynamic, and conserved r-loop structures associate with specific epigenomic signatures in mammals. *Molecular cell*, 63(1):167–178.

[Schmitz et al., 2010] Schmitz, K.-M., Mayer, C., Postepska, A., and Grummt, I. (2010). Interaction of noncoding rna with the rdna promoter mediates recruitment of dnmt3b and silencing of rrna genes. *Genes & development*, 24(20):2264–2269.

[Selby and Sancar, 1997] Selby, C. P. and Sancar, A. (1997). Cockayne syndrome group b protein enhances elongation by rna polymerase ii. *Proceedings of the National Academy of Sciences*, 94(21):11205–11209.

[Sollier et al., 2014] Sollier, J., Stork, C. T., García-Rubio, M. L., Paulsen, R. D., Aguilera, A., and Cimprich, K. A. (2014). Transcription-coupled nucleotide excision repair factors promote r-loop-induced genome instability. *Molecular cell*, 56(6):777–785.

[Stork et al., 2016] Stork, C. T., Bocek, M., Crossley, M. P., Sollier, J., Sanz, L. A., Chedin, F., Swigut, T., and Cimprich, K. A. (2016). Co-transcriptional r-loops are the main cause of estrogen-induced dna damage. *Elife*, 5:e17548.

[Tan-Wong et al., 2019] Tan-Wong, S. M., Dhir, S., and Proudfoot, N. J. (2019). R-loops promote antisense transcription across the mammalian genome. *Molecular cell*, 76(4):600–616.

[Teng et al., 2018] Teng, Y., Yadav, T., Duan, M., Tan, J., Xiang, Y., Gao, B., Xu, J., Liang, Z., Liu, Y., Nakajima, S., and Lan, L. (2018). Ros-induced r loops trigger a transcription-coupled but brca1/2-independent homologous recombination pathway through csb. *Nature communications*, 9(1):1–12.

[Thomas et al., 1976] Thomas, M., White, R. L., and Davis, R. W. (1976). Hybridization of rna to double-stranded dna: formation of r-loops. *Proceedings of the National Academy of Sciences*, 73(7):2294–2298.

[Vanoosthuyse, 2018] Vanoosthuyse, V. (2018). Strengths and weaknesses of the current strategies to map and characterize r-loops. *Non-coding RNA*, 4(2):9.

[Wahba et al., 2016] Wahba, L., Costantino, L., Tan, F. J., Zimmer, A., and Koshland, D. (2016). S1-drip-seq identifies high expression and polya tracts as major contributors to r-loop formation. *Genes & development*, 30(11):1327–1338.

[Wang et al., 2021] Wang, K., Wang, H., Li, C., Yin, Z., Xiao, R., Li, Q., Xiang, Y., Wang, W., Huang, J., Chen, L., and Liang, K. (2021). Genomic profiling of native r loops with a dna-rna hybrid recognition sensor. *Science Advances*, 7(8):eabe3516.

[Weinreb et al., 2021] Weinreb, J. T., Ghazale, N., Pradhan, K., Gupta, V., Potts, K. S., Tricomi, B., Daniels, N. J., Padgett, R. A., De Oliveira, S., and Verma, A. (2021). Excessive r-loops

trigger an inflammatory cascade leading to increased hspc production. *Developmental cell*, 56(5):627–640.

[White and Hogness, 1977] White, R. L. and Hogness, D. S. (1977). R loop mapping of the 18s and 28s sequences in the long and short repeating units of drosophila melanogaster rdna. *Cell*, 10(2):177–192.

[Wu et al., 2020] Wu, T., Lyu, R., You, Q., and He, C. (2020). Kethoxal-assisted single-stranded dna sequencing captures global transcription dynamics and enhancer activity in situ. *Nature methods*, 17(5):515–523.

[Wulfridge and Sarma, 2021] Wulfridge, P. and Sarma, K. (2021). A nuclease-and bisulfite-based strategy captures strand-specific r-loops genome-wide. *Elife*, 10:e65146.

[Xu et al., 2017a] Xu, J., Lahiri, I., Wang, W., Wier, A., Cianfrocco, M. A., Chong, J., Hare, A. A., Dervan, P. B., DiMaio, F., Leschziner, A. E., and Wang, D. (2017a). Structural basis for the initiation of eukaryotic transcription-coupled dna repair. *Nature*, 551(7682):653–657.

[Xu et al., 2020] Xu, J., Wang, W., Xu, L., Chen, J.-Y., Chong, J., Oh, J., Leschziner, A. E., Fu, X.-D., and Wang, D. (2020). Cockayne syndrome b protein acts as an atp-dependent processivity factor that helps rna polymerase ii overcome nucleosome barriers. *Proceedings of the National Academy of Sciences*, 117(41):25486–25493.

[Xu et al., 2017b] Xu, W., Xu, H., Li, K., Fan, Y., Liu, Y., Yang, X., and Sun, Q. (2017b). The r-loop is a common chromatin feature of the arabidopsis genome. *Nature plants*, 3(9):704–714.

[Yan and Sarma, 2020] Yan, Q. and Sarma, K. (2020). Mapr: A method for identifying native r-loops genome wide. *Current protocols in molecular biology*, 130(1):e113.

[Yasuhara et al., 2018] Yasuhara, T., Kato, R., Hagiwara, Y., Shiotani, B., Yamauchi, M., Nakada, S., Shibata, A., and Miyagawa, K. (2018). Human rad52 promotes xpg-mediated r-loop processing to initiate transcription-associated homologous recombination repair. *Cell*, 175(2):558–570.

[Yu et al., 2003] Yu, K., Chedin, F., Hsieh, C.-L., Wilson, T. E., and Lieber, M. R. (2003). R-loops at immunoglobulin class switch regions in the chromosomes of stimulated b cells. *Nature immunology*, 4(5):442–451.

[Zhang et al., 2014] Zhang, Z. Z., Pannunzio, N. R., Hsieh, C.-L., Yu, K., and Lieber, M. R. (2014). The role of g-density in switch region repeats for immunoglobulin class switch recombination. *Nucleic acids research*, 42(21):13186–13193.