GENOME-WIDE UV-INDUCED DNA DAMAGE AND NUCLEOTIDE EXCISION REPAIR IN THE CONTEXT OF R-LOOPS

by
SEZGI KAYA

Submitted to the Graduate School of
Engineering and Natural Sciences
in partial fulfilment of
the requirements for the degree of Doctor of Philosophy

Sabancı University
December 2023

EXPLORING THE UV-INDUCED DNA DAMAGE AND NUCLEOTIDE EXCISION REPAIR IN THE CONTEXT OF R-LOOPS

Approved by:

.....

(Dissertation Supervisor)

.....

.....

.....

.....

Date of Approval: December 28, 2023

ABSTRACT

GENOME-WIDE UV-INDUCED DNA DAMAGE AND NUCLEOTIDE EXCISION REPAIR IN THE CONTEXT OF R-LOOPS

SEZGI KAYA

Molecular Biology, Genetics and Bioengineering Ph.D DISSERTATION, DECEMBER 2023

Dissertation Supervisor: Asst. Prof. Ogün Adebali

Keywords: R-loop, UV, NER, CPD, ssDNA, DNA:RNA hybrid

R-loops are three-stranded nucleic acid structures formed frequently during transcription when nascent RNA anneals on its complementary DNA strand, leaving the other DNA strand single-stranded. R-loops are dynamic structures that are formed and resolved by a number of regulator proteins such as helicases and endonucleases. Under this tight regulation, they play roles in important cellular processes such as gene expression regulation, transcription termination, immunoglobulin class switch recombination, telomere maintenance. However, upon accumulation, R-loops lead to double-strand breaks (DSBs) and genome instability.

UV exposure leads to formation of bulky lesions on DNA by triggering a photochemical reaction that results in covalent bonds between adjacent pyrimidines. These bulky lesions are cyclobutane pyrimidine dimers (CPDs) and (6-4) pyrimidine pyrimidone photoproducts ((6-4)PP) which might turn into mutations if not repaired. Nucleotide excision repair (NER) is the main mechanism for the repair of UV-induced lesions. To date, there is no clear evidence on how R-loops affect UV-induced damage formation and their repair by NER mechanism. Here, a comprehensive look at the relationship between R-loops and UV-induced damage formation and repair in human and *Arabidopsis* genomes is presented. Firstly, the R-loop locations determined by different methods and databases were compared in terms of genomic distribution and chromatin features, and the most reliable set of R-loops were selected for further analysis. Secondly, the aspects of UV-induced damage formation on R-loop are examined and the differences in CPD damage accu-

mulation patterns between R-loop strands and other genomic regions are addressed with molecular dynamics (MD) simulations. Then, the efficiency of UV-induced damage repair is assessed comparatively between R-loop strands and other genomics regions. Additionally, repair efficiency on R-loops is examined from another perspective, mutational burden. To do this, the distribution of mutations on the genomes of melanoma patients on R-loops and other genomic regions is presented in comparison to other cancer genomes. In the final part, Hidden Markov Models (HMMs) is used to classify human genome into states considering the differential occupancy of R-loop regulatory proteins, and these states are compared in terms of damage formation and repair efficiency. Findings of this study provide a broadened sight for R-loop - NER relationship.

ÖZET

GENOME-WIDE UV-INDUCED DNA DAMAGE AND NUCLEOTIDE EXCISION REPAIR IN THE CONTEXT OF R-LOOPS

SEZGİ KAYA

Moleküler Biyoloji, Genetik and Biyomühendislik Doktora Tezi, Aralık 2023

Tez Danışmanı: Asst. Prof. Ogün Adebali

Anahtar Kelimeler: R-loop, UV, NER, CPD, ssDNA, DNA:RNA hybrid

R-loop'lar transkripsiyon sırasında sıkça meydana gelen nükleik asit yapılarıdır. Bu yapılar, yeni oluşan RNA'nın komplementer DNA zinciri üzerine bağlanmasıyla, karşı DNA zincirini tek iplikli bırakarak oluşurlar. R-loop'lar, helikazlar ve endonükleazlar gibi bir dizi düzenleyici protein tarafından çözülen dinamik yapılar olarak sıkı bir kontrol altındadırlar. Gen ifadesi düzenlemesi, transkripsiyon sonlandırma, immünglobulin sınıf değişim rekombinasyonu, telomer bakımı gibi önemli hücresel süreçlerde rol oynarlar. Ancak R-loop'lar, birikmeleri durumunda, çift zincir kırılmalarına ve genom bütünlüğünün bozulmasına neden olabilirler.

UV maruziyeti, bitişik pirimidinler arasında kovalent bağlara yol açan fotokimyasal bir reaksiyonu tetikleyerek DNA üzerinde lezyonların oluşmasına neden olur. Bu lezyonlar, siklobütan pirimidin dimerleri (CPD'ler) ve (6-4) pirimidin pirimidon fotoürünlerinin ((6-4)PP'ler) olarak adlandırılır ve onarılmazlarsa mutasyonlara dönüşebilir. Nükleotid nükleotid eksizyon onarımı (NER), UV tarafından oluşturulan lezyonların onarımı için ana mekanizmadır. Şu ana kadar, R-loop'ların UV kaynaklı hasar oluşumuna ve bu hasarların NER mekanizmasıyla onarımına nasıl etki ettiğine dair net bir kanıt bulunmamaktadır. Bu çalışmada, insan ve *Arabidopsis* genomlarındaki UV tarafından oluşturulan hasarlar ile R-loop'lar arasındaki ilişkiye kapsamlı bir bakış sunulmaktadır. İlk olarak, genomda R-loop bölgelerindeki CPD hasarı oluşumu incelenmiş ve R-loop zincirleri ile diğer genomik bölgeler arasındaki hasar birikim oranlarının farklılıkları moleküler dinamik (MD) simülasyonlarıyla ele alınmıştır. İkinci olarak, CPD onarım verimliliği, R-loop zincirleri ile diğer genomik bölgeler arasında karşılaştırmalı olarak değerlendirilmiştir.

Ardından, R-loop'larda onarım verimliliği, mutasyon yükü açısından incelenmiştir. Bunun için, melanoma hastalarının genomları üzerindeki mutasyonların R-loop'lar ve diğer genomik bölgeler arasındaki farklı dağılımı analiz edilmiştir. Son bölümde, Hidden Markov Modelleri (HMM'ler), R-loop'ları regüle eden proteinlerin genomdaki farklı bağlanma durumları göz önüne alınarak insan genomunu sınıflandırmak için kullanılmış ve bu sınıflarda bulunan R-loop'lar hasar oluşumu ve onarım verimliliği açısından karşılaştırılmıştır. Bu çalışmanın bulguları, R-loop - NER ilişkisine geniş bir bakış sağlamaktadır.

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my advisor Asst. Prof. Dr. Ogün Adebali, for their unwavering support and guidance, and endless patience throughout the journey of my Ph.D. Their expertise and mentorship have been invaluable in shaping the direction of my research and academic growth.

I am also thankful to my committee members, Prof. Dr. Levent Öztürk, Asst. Prof. Dr. Onur Öztaş, Prof. Dr. Canan Atılgan and Asst. Prof. Dr. Emrah Nikerel, for their insightful feedback and constructive criticism, which greatly enriched the quality of this thesis.

My sincere appreciation goes to Cem Azgari, Yağmur Sözeri Güneri, Aylin Bircan, Veysel Oğulcan Kaya, Nurdan Kuru, Ümit Akköse, Mustafa Malkoç, Arda Temena and Berkay Selçuk from Adebali Lab for their invaluable support, collaborative spirit, and shared experiences, which made this journey exceptional. I would like to extend my sincere thanks to Ayesha Fatima, Tuba Sena Oğurlu and Baran Özcan for their invaluable contribution that have greatly enriched my research.

I express profound gratitude to my loving husband, and my affectionate family and friends for their consistent support, love, and understanding during this journey. Your belief in me has been my motivation to keep going. A very special thanks to Pesto and all other furry friends for all the joy and emotional support.

I would like to acknowledge the financial support provided by TÜBİTAK, which made this research possible.

To my loving husband Onur
and to all my family

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

UV ultraviolet

NER nucleotide excision repair

GG-NER global genome nucleotide excision repair

TC-NER transcription-coupled nucleotide excision repair

CPD cyclobutane pyrimidine dimers

(6-4)PP (6-4) pyrimidine pyrimidone photoproducts

ssDNA single-strand DNA

DSB double-strand break

TS transcribed strand

NTS non-transcribed strand

HMM Hidden Markov Models

MD molecular dynamics

1. INTRODUCTION

1.1 R-loops as both beneficial and harmful structures

R-loops are three-stranded structures that occur on DNA when an RNA molecule binds to its complementary DNA strand, resulting in the other DNA strand being left as single-stranded DNA (ssDNA) (Figure 1.1). They are mostly formed during transcription when the nascent RNA binds to its complementary DNA strand in close proximity to its transcription site, forming cis R-loops. Trans R-loops also exist which form when an RNA molecule binds to a complementary DNA strand far from its transcription site (Sollier & Cimprich, 2015). R-loops are abundant structures found on the genomes of many organisms, from bacteria to humans and they form naturally in normal conditions (Sollier & Cimprich, 2015). They are present in 5% of mammalian genomes and 10% of the *Arabidopsis* genome (Sanz, Hartono, Lim, Steyaert, Rajpurkar, Ginno, Xu & Chédin, 2016; Xu, Xu, Li, Fan, Liu, Yang & Sun, 2017). So far, R-loops have been linked to several processes such as transcription, safeguarding promoters from methylation to regulate gene expression, immunoglobulin class switch recombination, double-strand break (DSB) repair (Costantino & Koshland, 2015; Gómez-González & Aguilera, 2020). On the other hand, previous studies have found that R-loops lead to DSBs when not resolved with the necessary timing (Costantino & Koshland, 2018; Crossley, Bocek & Cimprich, 2019; Sollier & Cimprich, 2015).

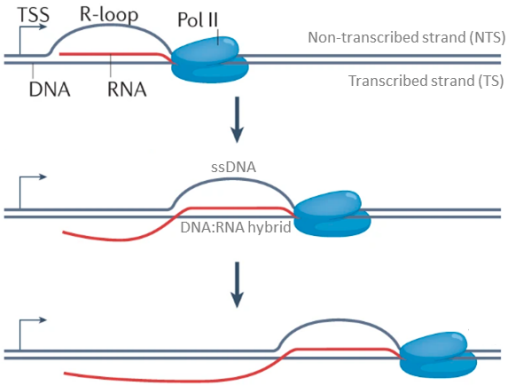

Figure 1.1 Formation of R-loop structures during transcription (Petermann, Lan & Zou, 2022).

1.1.1 Roles of R-loops in cellular processes

R-loops have been associated with several cellular processes in previous studies, such as transcription regulation (Niehrs & Luke, 2020). In *Arabidopsis*, the expression of long non-coding RNA (lncRNA) *COOLAIR* is silenced by the formation of an R-loop on the promoter region (Sun, Csorba, Skourti-Stathaki, Proudfoot & Dean, 2013). Another example is the *GAL* gene cluster-associated lncRNAs which promote gene expression by forming R-loops in yeast (Cloutier, Wang, Ma, Al Husini, Dhoondia, Ansari, Pascuzzi & Tran, 2016; Hegazy, Fernando & Tran, 2020).

In addition to the functions associated to lncRNAs, R-loops protect promoters from DNA methylation and thus, promote gene expression (Grunseich, Wang, Watts, Burdick, Guber, Zhu, Bruzel, Lanman, Chen, Schindler & others, 2018). They also serve as scaffolds for transcription factors (TFs). At the transcription start site of the *VIM* gene, R-loop formation recruits nuclear factor-KB (NF-KB) and induce transcription (Niehrs & Luke, 2020). R-loops also regulate gene expression by modulating chromatin accessibility. For instance, R-loop presence increased the levels of histone H3 Ser-10 phosphorylation (H3S10P) leading to the condensation of the chromatin in yeast. The same study stated that the RNase H overexpression caused a decrease in H3S10P, indicating that R-loops are related to the formation

of closed chromatin (Castellano-Pozo, Santos-Pereira, Rondón, Barroso, Andújar, Pérez-Alegre, García-Muse & Aguilera, 2013; Hegazy et al., 2020). It was also found that the resolution of the R-loop formed by the *COOLAIR* lncRNA resulted in chromatin silencing via recruitment of chromatin modifier proteins (Xu, Wu, Duan, Fang, Jia & Dean, 2021).

R-loops tend to form at the CpG islands of gene promoters where the GC skew is favorable for R-loop formation. Another way that R-loops use to impact gene expression is protecting promoters from methylation (Al-Hadid & Yang, 2016). On the promoter *BAMBI* gene, R-loops prevented methyltransferase DNMT1 from methylating the promoter and silencing the gene. When RNase H was overexpressed, the binding of DNMT1 on the promoter was more frequent, suggesting that the R-loop presence was the factor making DNMT1 binding more difficult (Grunseich et al., 2018). R-loops also promote demethylation on the promoters. *TCF21* gene expression is induced by GADD45A protein which binds and recruits TET1 for demethylation. In addition, the study suggested that *TCF21* was not the only gene on the genome that was regulated by the TET1 binding on R-loops. They identified numerous CpG island with R-loops that were the targets for TET1 binding (Arab, Karaulanov, Musheev, Trnka, Schäfer, Grummt & Niehrs, 2019).

Termination of transcription is an important process since unterminated RNA polymerases can also transcribe the downstream genes and interfere with the sensitive balance of gene expression regulation (Niehrs & Luke, 2020). To efficiently stop transcription, R-loops are located at the 3' ends of many genes to provide a physical obstacle for the RNA polymerases (Ginno, Lim, Lott, Korf & Chédin, 2013). Another study identified m6A RNA modification as the inducer of R-loop formation on gene terminals. It was found that the loss of m6A methyltransferase METTL3 resulted in R-loop reduction on transcription end sites (TES) and impaired transcription termination (Yang, Liu, Xu, Zhang, Yang, Ju, Chen, Chen, Li, Ren & others, 2019).

Another function of R-loops was discovered in DSB repair. R-loop accumulation was detected upon DSB formation in human and yeast genomes (Lu, Hawley, Skalka, Baldock, Smith, Bader, Malewicz, Watts, Wilczynska & Bushell, 2018; Ohle, Tesorero, Schermann, Dobrev, Sinning & Fischer, 2016). In humans, this accumulation was specific to DSB sites since no significant change was detected on the other parts of the genome by S9.6 antibody. Overexpression of RNase H1 resulted in diminished efficiency of DSB repair both by homologous recombination (HR) and nonhomologous end joining (NHEJ) pathways in humans and by HR in yeast, indicating that R-loop formation on DSB is important for the maintanance of the

genome (Lu et al., 2018; Ohle et al., 2016).

Finally, a function of R-loops that was discovered the earliest is in class switch recombination (CSR) in B cells. CSR is a recombination process in vertebrates that changes the constant region of the heavy chain which results in the change of immunoglobulin type from IgM any other type (Yu, Chedin, Hsieh, Wilson & Lieber, 2003). The ssDNA strand of the R-loops was found to attract AID protein which is necessary for CSR. Reduced R-loop levels led to CSR impairment indicating the necessity to R-loops in CSR efficiency (Wiedemann, Peycheva & Pavri, 2016; Zhang, Pannunzio, Han, Hsieh, Yu & Lieber, 2014).

1.1.2 R-loops as threats to genome integrity

In spite of contributing to many cellular pathways, accumulation of unresolved R-loops may cause DNA damage leading to impaired genome integrity and cancer (Sollier & Cimprich, 2015). One of the ways that R-loops cause DNA damage is interfering with the replication fork (Crossley et al., 2019). R-loops were found inducing transcription-replication collisions (TRCs) under the estrogen treatment condition whereas RNase H overexpression reduced this outcome (Stork, Bocek, Crossley, Sollier, Sanz, Chedin, Swigut & Cimprich, 2016). In addition, at the S phase, R-loop accumulation affected replication by slowing down the fork progression while this effect was restored by RNase H overexpression (Brickner, Garzon & Cimprich, 2022). R-loops also caused fork stalling and the induction of DNA recombination while this phenotype was also returned by RNase H overexpression, suggesting that under the conditions which favor R-loop accumulation, R-loops can be lethal by leading to replication stress (Gan, Guan, Liu, Gui, Shen, Manley & Li, 2011; Hegazy et al., 2020).

In addition to replication fork, unresolved R-loops may also lead to RNA polymerase stalling during transcription (Rinaldi, Pizzul, Longhese & Bonetti, 2021). Notably, the reverse of this effect was also detected where the stalled RNA polymerases induce R-loop formation from the nascent RNA that had been produced (Shivji, Renaudin, Williams & Venkitaraman, 2018). Normally, the DNA lesions that stall RNA polymerase induce transcription-coupled nucleotide excision repair (TC-NER) machinery which lead to the excision of the damaged DNA strand by XPF and XPG endonucleases (Reardon & Sancar, 2005). Previous studies found that XPF and XPG also cut the accumulated R-loops which led to DSB formation although the reason and the mechanism behind this process are not clear (Sollier, Stork,

García-Rubio, Paulsen, Aguilera & Cimprich, 2014).

Another factor that makes R-loops threats to genome integrity is the ssDNA that is present on one of the strands of the R-loops. ssDNAs are generally more prone to DNA damaging agents that lead to mutations and DNA breaks (Rinaldi et al., 2021). For instance, AID protein naturally turn cytosine nucleotide to uracils which are then processed by the base excision repair (BER) machinery leading to nicks on the DNA and eventually, DSBs (So & Martin, 2019; Stavnezer & Schrader, 2006; Yu et al., 2003). Moreover, reactive oxygen species (ROS) can also create damages on the vulnerable ssDNA (Brickner et al., 2022). RNA-modifiying enzymes such as ADAR1 can also create nicks on the DNA strand within the DNA:RNA hybrid which may lead to single-strand breaks (SSBs) (Zheng, Lorenzo & Beal, 2017).

1.1.3 Regulation of R-loops

Accumulation of R-loops may cause breaks on the genome causing lethal effects besides their beneficial roles in many cellular processes. Thus, the regulation of these structures is a crucial task for the cells. RNase H is a well-known R-loop regulator which dissolves the RNA moiety in the DNA:RNA hybrid strand of R-loops (Zhao, Zhu, Limbo & Russell, 2018). RNase H1 and RNase H2 are both conserved in many organisms. RNase H2 works at certain times of the cell cycle while RNase H1 regulation is independent of the cell cycle and responsive R-loop-induced stress (Lockhart, Pires, Bento, Kellner, Luke-Glaser, Yakoub, Ulrich & Luke, 2019). RNase H is used in numerous experiments to confirm various effects of R-loops (Mackay, Xu & Weinberger, 2020). In addition, a mutated version of RNase H that can bind to R-loops but cannot resolve them is used in methods that sequence R-loop positions on the genome (Tan-Wong, Dhir & Proudfoot, 2019). Another nuclease, exoribonuclease 2 (XRN2), which normally cleaves the 3' end of the nascent RNA to support transcriptional termination, was also detected as and R-loop regulator whose deficiency resulted in R-loop accumulation (Morales, Richard, Patidar, Motea, Dang, Manley & Boothman, 2016; Villarreal, Mersaoui, Yu, Masson & Richard, 2020).

Besides nucleases, cells also use DNA:RNA helicases to resolve R-loop structures. Senataxin (SETX) is one of the well-identified DNA:RNA helicases that was found in human and yeast cells (Gatti, De Domenico, Melino & Peschiaroli, 2023). DHX9 is another helicase that dissociates RNA from DNA on R-loops (Chakraborty & Grosse, 2011) even though it also induced R-loop formation under specific conditions

(Chakraborty, Huang & Hiom, 2018). DDX21 which is also a helicase resolve R-loops in coordination with SIRT7 (Song, Hotz-Wagenblatt, Voit & Grummt, 2017).

Numerous other proteins also involve in R-loop resolution by recruiting the necessary factor on the R-loop sites. BRCA1 and BRCA2 are also known to regulate R-loops (Hatchi, Skourti-Stathaki, Ventz, Pinello, Yen, Kamieniarz-Gdula, Dimitrov, Pathania, McKinney, Eaton & others, 2015; Shivji et al., 2018). BRCA1 recruits SETX helicase to 3' ends of the genes to resolve R-loops. While BRCA1 was detected in the regulation of both 5' and 3' R-loops, BRCA2 regulated 5' R-loops only. Together, these two proteins prevent RNA polymerase stalling and transcription stress by resolving R-loops that may interfere with the RNA polymerase (Hatchi et al., 2015; Shivji et al., 2018; Zhang, Chiang, Wang, Zhang, Smith, Zhao, Nair, Michalek, Jatoi, Lautner & others, 2017). RPA is another protein that recruits RNase H1. It binds to ssDNA strands of R-loops and helps maintain genome integrity (Nguyen, Yadav, Giri, Saez, Graubert & Zou, 2017).

1.1.4 Methods to map R-loops

The literature on R-loops holds a set of methods that use next-generation sequencing to determine R-loop positions on the genome. These methods, in general, use immonuprecipitation technique which depend on either of the two molecules: (1) S9.6 antibody, (2) mutated RNase H protein (Figure 1.2). S9.6 antibody targets DNA:RNA hybrids even though its specificity is questioned due to having affinity to double-stranded RNAs (dsRNAs) at some level (Bou-Nader, Bothra, Garboczi, Leppla & Zhang, 2022; Hartono, Malapert, Legros, Bernard, Chédin & Vanoosthuyse, 2018). Mutated RNase specifically binds DNA:RNA hybrids while it was shown that RNase H can also bind to RPA-bound ssDNA for R-loop resolution (Nguyen et al., 2017).

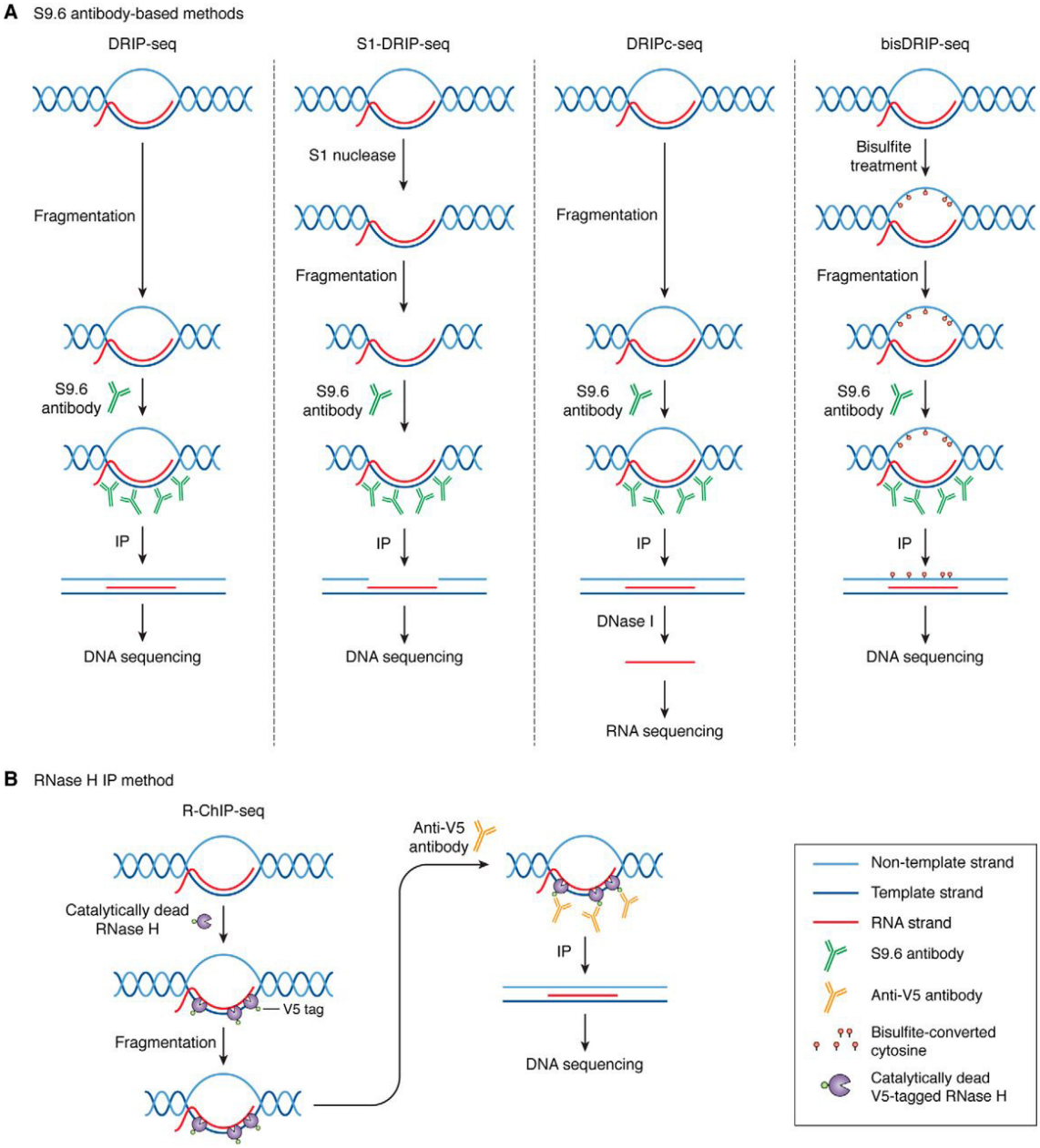

Figure 1.2 Methods to map R-loops on genome (Hegazy et al., 2020). (A) Methods based on S9.6 antibody to pull down the R-loops. (B) Methods based on mutant RNase H1 which lacks its nuclease function.

Methods that use S9.6 antibody-based pull down of R-loops are termed as DNA:RNA immunoprecipitation sequencing (DRIP-seq). Along with the DRIP-seq method (Hamperl, Bocek, Saldivar, Swigut & Cimprich, 2017), variations of it were also adopted. ssDRIP-seq method was developed to add strand specificity to DRIP-seq by modifiying the library preparation step (Xu et al., 2017; Yang et al., 2019). qDRIP-seq method sequenced DNA moiety in the DNA:RNA hybrid with the integration of synthetic hybrids in order to perform a better normalization between the samples (Crossley, Bocek, Hamperl, Swigut & Cimprich, 2020). DRIPc-seq cap-

tured the RNA moiety of the DNA:RNA hybrid to increase the resolution (Sanz et al., 2016). Finally, bisDRIP-seq aimed to increase the resolution by marking the cytosines on ssDNA with deamination by sodium bisulfite. S9.6 antibody is then used to capture the R-loops (Dumelie & Jaffrey, 2017).

To overcome the specificity issues associated with the S9.6 antibody, RNase H-based methods were developed. In general, the catalytic site of the RNase H, which normally degrades the RNA moiety in DNA:RNA hybrids, is inactivated by D210N mutation. In this technique, another version of RNase H1 (WKKD) is also produced which carries four mutations to inactivate both the catalytic and the binding domains in order to be used as negative control (Chen, Chen, Zhang, Gu, Xiao, Shao, Tang, Qian, Luo, Li & others, 2017; Tan-Wong et al., 2019). The two examples of this approach are R-ChIP-seq (Chen et al., 2017) and RR-ChIP-seq (Tan-Wong et al., 2019). Both methods sequenced the RNA within the DNA:RNA hybrid of R-loops while RR-ChIP-seq additionally overexpressed the mutant RNase H1 and treated samples with DNase before sequencing (Tan-Wong et al., 2019).

1.2 UV-induced damage formation on DNA

Exposure to ultraviolet (UV) light causes bulky lesions on DNA. These lesions are of two subtypes: cyclobutane pyrimidine dimers (CPD) and pyrimidine (6-4) pyrimidone photoproducts ((6-4)PP). CPD is the most common type of these lesions, which is formed through a [2+2] cycloaddition process between the C5–C6 double bonds of two pyrimidine bases (Cadet, Mouret, Ravanat & Douki, 2012). CPD lesions are more commonly result in mutations due to being formed more frequently and repaired more slowly. A significant proportion of mutations caused by UVB radiation are C to T transitions specifically occurring at dipyrimidine sites, which is the prevalent type of mutation in skin cancer genomes (Leffell, 2000; Pfeifer, 2020).

CPDs can form between two adjacent pyrimidines although the thymine-thymine (TT) positions provide the most favorable conditions for the reaction (Lee & Matsika, 2022). Despite this reactivity, the positions of the adjacent pyrimidines relative to each other at the time of UV exposure is the most important criteria for CPD formation to be finalized since unfavorable positions may change due to the movement of the DNA backbone (Law, Azadi, Crespo-Hernández, Olmon & Kohler, 2008). Two components of this relative positioning is considered crucial for cycloaddition

reaction to occur: the distance and the dihedral angle between C5-C6 double bonds of the adjacent Ts. As the distance and angle increase, the probability of the reaction decreases (Law et al., 2008; Nayis, Liebl & Zacharias, 2023; Stark, Poon & Wyrick, 2022). In addition to the distance and the angle, studies have tested the impact of the flanking nucleotides on both sides of the TT pairs and stated that the reactivity of the TTs change depending on the nucleotides adjacent to them (Law, Forties, Liu, Poirier & Kohler, 2013; Lu, Gutierrez-Bayona & Taylor, 2021; Pan, Hariharan, Arkin, Jalilov, McCullagh, Schatz & Lewis, 2011).

6-4PPs are formed through a more gradual two-step process, which involves the formation of a UV-induced oxetane (or azetidine) intermediate (Bohm, Morledge-Hampton, Stevison, Mao, Roberts & Wyrick, 2023). Thymine-thymine (TT) and thymine-cytosine (TC) pairs are the most common dipyrimidines for (6-4)PPs formation (Hu, Adebali, Adar & Sancar, 2017). (6-4)PPs induce a more significant alteration in the DNA conformation compared to CPDs. This suggests that (6-4)PPs may not be easily accommodated or formed within the tightly coiled nucleosomal DNA (Mao, Wyrick, Roberts & Smerdon, 2017). In addition, both CPD and (6-4)PP damage formations are reported as being affected from the binding of transcription factors (Aboussekhra & Thoma, 1999; Frigola, Sabarinathan, Gonzalez-Perez & Lopez-Bigas, 2021; Mao, Brown, Esaki, Lockwood, Poon, Smerdon, Roberts & Wyrick, 2018; Tornaletti & Pfeifer, 1995).

Several methods have been developed to map UV-induced damages on the genome upon UV exposure. These methods include Excision-seq (Bryan, Ransom, Adane, York & Hesselberth, 2014), CPD-seq (Mao & Wyrick, 2020) and Damage-seq (Hu et al., 2017; Hu, Lieb, Sancar & Adar, 2016). Despite being developed for the same purpose, Damage-seq became a better option due to its higher sensitivity and requirement of lower UV dosage (Figure 1.3). It captures the damaged positions on the genome using damage-specific antibodies followed by a strand-specific library preparation (Hu et al., 2017).

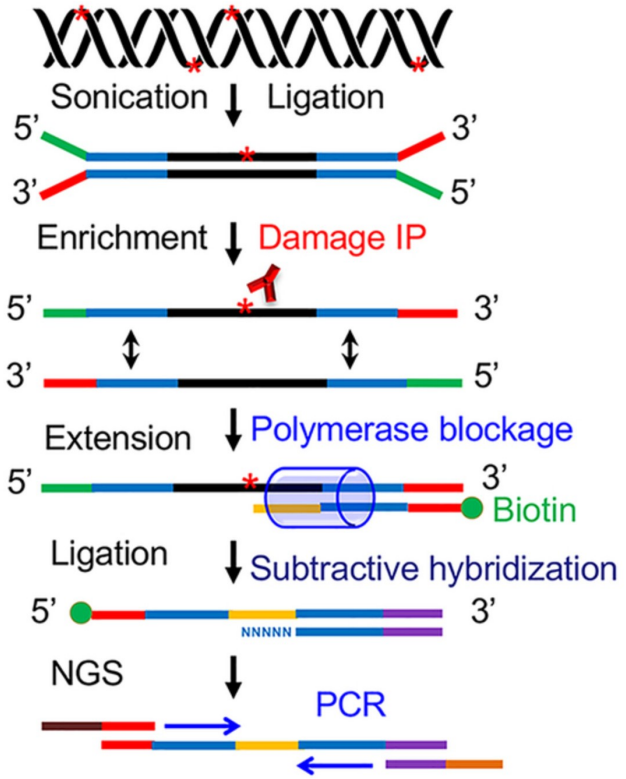

Figure 1.3 Schematic representation of Damage-seq method (Li & Sancar, 2020).

The result is a transition of C to T and CC to TT [4], which are the most frequent mutations of p53 in both human and mouse skin cancers

mutations cancer

1.3 Nucleotide excision repair

Cells encounter numerous types of DNA damaging agents throughout their lives and huge amounts of damage forms on the DNA every day (Marteijn, Lans, Vermeulen & Hoeijmakers, 2014). Since it is not possible to change the damaged DNA with an undamaged one, the integrity of the genome solely depends on its efficient and prompt repair. The presence of DNA damages at the time of replication may lead to mutations and thus, cancer (Reardon & Sancar, 2005). To protect genomes from these damages and maintain the genome integrity, multiple mechanisms have been developed evolutionarily to efficiently deal with different types of damages. These mechanisms are parts of an overall DNA damage response (DDR) which includes

elements to detect the damage and start signaling to activate the corresponding repair mechanism and eventually, determine the fate of the cell (Marteijn et al., 2014).

Ultraviolet (UV) light is one of the most common damaging agents and nucleotide excision repair (NER) is the main mechanism evolved to cope with UV-induced DNA damages (Reardon & Sancar, 2005). UV-induced damages are not the only targets for NER; other bulky lesions and bondings between the nucleotides of the same DNA strand caused by chemicals and drugs, and cyclopurines induced by ROS are also repaired by NER mechanism (Reardon & Sancar, 2005). Defects in NER have been linked to various autosomal recessive genetic illnesses in humans, including xeroderma pigmentosum (XP). Individuals afflicted with XP display a heightened sensitivity to sunlight and a notable susceptibility to developing skin cancer (Schärer, 2013; Sugasawa, 2008).

NER mechanism is composed of three steps (Figure 1.4). In the first step, damage is recognized by a set of proteins. There two subpathways of NER that differentiate in their damage recognition mechanisms: transcription-coupled nucleotide excision repair (TC-NER) and global genome nucleotide excision repair (GG-NER). GG-NER can repair damages all around the genome while TC-NER preferentially repairs the actively transcribed regions (Schärer, 2013). In GG-NER, XPC protein searches and binds to the helix distortions on DNA caused by the damage (de Laat, Jaspers & Hoeijmakers, 1999a; Schärer, 2013). However, XPC cannot detect CPD damages since they do not create a large distortion on the DNA. For CPD recognition, UV-DDB complex which includes DDB1 and DDB2 binds on the damaged sites and bends the DNA in order for XPC to detect and bind (Scrima, Koníčková, Czyzewski, Kawasaki, Jeffrey, Groisman, Nakatani, Iwai, Pavletich & Thomä, 2008). On the other hand, damage recognition depends on the stalled RNA polymerase at the damaged site in TC-NER. Then, additional factors such as CSA, CSB and XAB2 are recruited to the damaged site (Schärer, 2013).

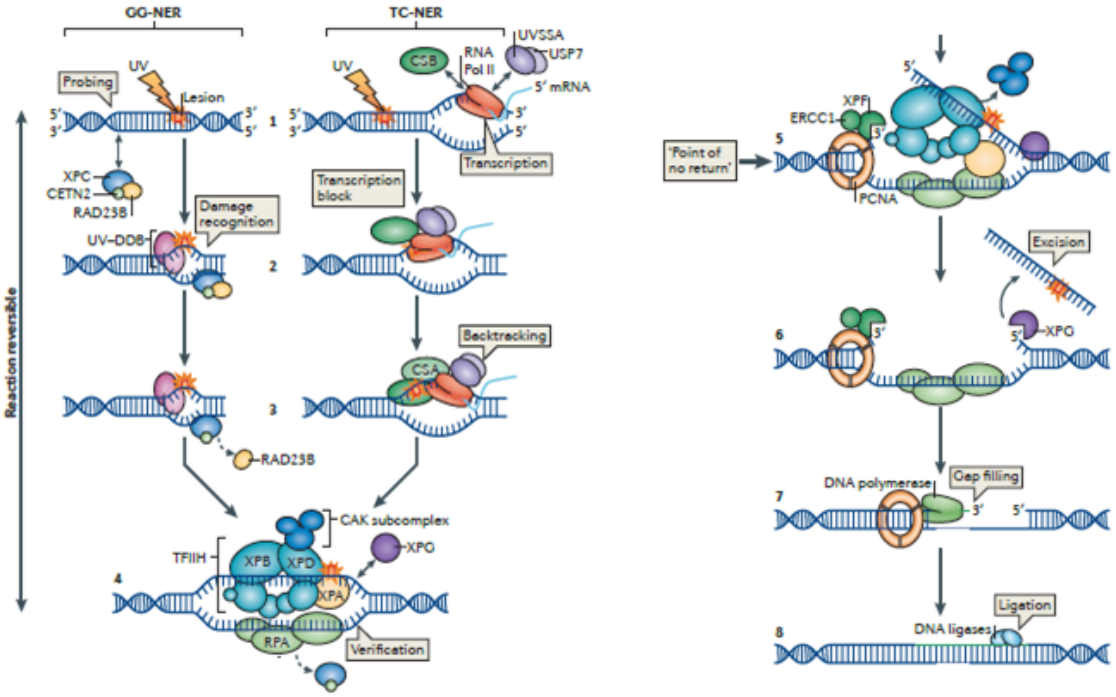

Figure 1.4 The mechanism of nucleotide excision repair (NER) (Marteijn et al., 2014).

After the recognition of the damage, GG-NER and TC-NER subpathways follow the same mechanism. The TFIIH complex assembles at the damage site which contains two helicases, XPB and XPD, which open up the two strands of the DNA for the repair (Compe & Egly, 2012). XPD also performs damage verification and stops the repair process if it does not encounter the damage (Mathieu, Kaczmarek, Rüthemann, Luch & Naegeli, 2013). After the complete assembly of TFIIH complex, XPG, XPA and RPA proteins are recruited to the damage site. XPG is the endonuclease that excises the 3' of the damage site (Marteijn et al., 2014). XPA is a mediator protein which interacts with almost all NER proteins. It also recruits the second endonuclease, XPF, for the incision of the 5' of the damage (Marteijn et al., 2014). The recruitment of XPF is a prerequisite for the excision process to start (Schärer, 2013). RPA binds on the DNA strand that is not damaged and coordinates the positioning of the XPG and XPF (De Laat, Appeldoorn, Sugasawa, Weterings, Jaspers & Hoeijmakers, 1998). When XPG and XPF are positioned accordingly, the excision of the damaged DNA is triggered. The excised DNA with the damage, or the excision product, is released from the DNA with the TFIIH complex that is bound to it (Kemp, Reardon, Lindsey-Boltz & Sancar, 2012), leaving a single-strand of 22-30 nucleotide length (Marteijn et al., 2014). PCNA ring is bound to the site immediately after the 5' excision. RCF protein and different types of DNA polymerases fill the gap left after the excision. Finally, DNA ligase 1 and 3 seal the nick

at the end of the newly synthesized DNA (Marteijn et al., 2014; Schärer, 2013).

NER efficiency can be altered by many factors in the cell. For instance, the low accessibility of the damaged site reduces the possibility of NER proteins to reach the damage (Polo & Almouzni, 2015). The presence of epigenetic modifications on the chromatin also affects NER while they also regulate DDR (Gong, Kwon & Smerdon, 2005; Marteijn et al., 2014). XR-seq (excision repair sequencing) has been developed to assess the NER activity genome-wide (Figure 1.5) (Hu, Adar, Selby, Lieb & Sancar, 2015). It also allowed us to compare the repair activities on distinct regions on the genome, as well as on the two strands of the DNA due to its strand-specificity. The method captures the released excised oligonucleotides with antibodies specific to damage types and it was able to map both CPDs and (6-4)PPs (Hu et al., 2015,1).

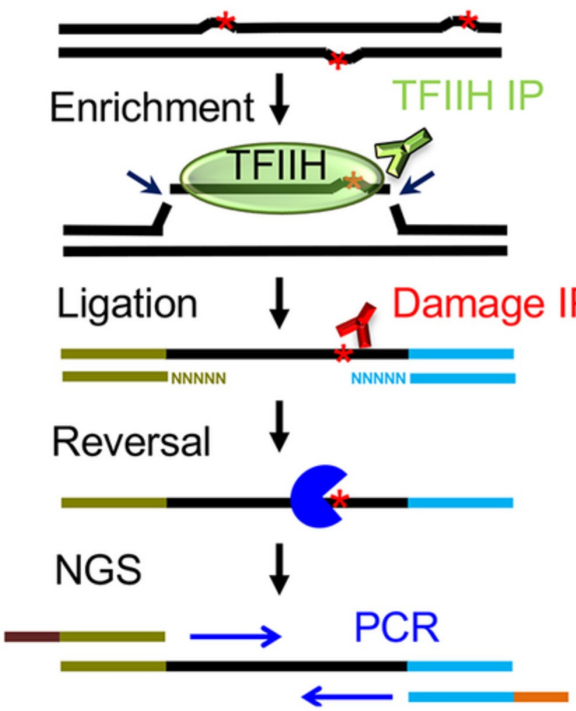

Figure 1.5 Schematic representation of XR-seq method (Li & Sancar, 2020).

2. THE SCOPE OF THE THESIS

R-loops are in the spotlight for years due to their expanding roles in many important cellular processes such as gene expression regulation and double-strand break (DSB) repair. On the other hand, they are defined as threats to genome integrity since excessive amounts of unresolved R-loops cause DSB formation. R-loops are prevalent three-stranded structures that naturally occur in our genomes during transcription when the nascent RNA anneals on the complementary DNA strand leaving the other DNA strand as single-stranded (ssDNA). However, to date, there is no clear information about how R-loop regions are affected from UV-induced damages or how efficiently those damages are repaired by nucleotide excision repair (NER) mechanism. Therefore, we aimed to shed light on R-loop-NER relationship to learn more about how these prevalent structures are maintained.

This thesis presents a detailed examination of the connection between R-loops and UV-induced damage formation and the repair efficiency in the human and *Arabidopsis* genomes. Firstly, we focused on processing R-loop data in the literature that were obtained using different methods. We compared them in terms of genome coverage, genomic distribution on different chromatin states, histone marker abundance to assess their accuracies and decided which set of R-loop positions to include in the further steps of our work.

Secondly, we focused on investigating the UV-induced damage formation on R-loop sites in comparison to the other sites in the genome. We used Damage-seq datasets to map the damaged sites on the genome. We specifically examined the differences in patterns of CPD damage accumulation between R-loop strands and other sections of the genome using molecular dynamics (MD) simulations. Furthermore, the comparative assessment of nucleotide excision repair (NER) efficiency is conducted across R-loop strands and other areas of the genome. The reasons for differential repair efficiency on R-loops was investigated by including time-point XR-seq data representing the two subpathways of NER, global genome NER (GG-NER) and transcription-coupled NER (TC-NER), as well as an XR-seq data obtained from TC-NER-inhibited cells. The repair efficiency of *Arabidopsis* genes and R-loops

were also examined.

Next, we shifted the focus to the mutational burden to assess repair efficiency on R-loops from another perspective. The distribution of mutations on the genomes of melanoma patients, specifically on R-loops and other genomic areas, is examined and compared. Expected mutations due to the sequence content were calculated and compared with the observed mutation counts in order to make a connection between the mutation events and repair efficiency.

Finally, Hidden Markov Models (HMMs) are employed to categorize the human genome into states based on the varying binding of R-loop regulator proteins. The R-loops from these states are compared in terms of chromatin accessibility, damage distribution and repair efficiency to check whether there are differing behaviors on R-loop subsets with different regulators. The findings of this study offer an expanded perspective on the interplay between R-loop and NER.

3. METHODS

3.1 Datasets used in this study

Human R-loop sites were retrieved from DRIP-seq (Hamperl et al., 2017), ssDRIP-seq (Yang et al., 2019), qDRIP-seq (Crossley et al., 2020) and RR-ChIP-seq (Tan-Wong et al., 2019) studies and R-loopBase (level 7 subset) (Lin, Zhong, Zhou, Geng, Hu, Huang, Hu, Fu, Chen & Chen, 2022) and RLBase (Miller, Montemayor, Abdul, Vines, Levy, Hartono, Sharma, Frost, Chédin & Bishop, 2022) databases. *Arabidopsis* R-loop sites were obtained from an ssDRIP-seq study (Xu et al., 2017).

Human NHF1 (Hu et al., 2017) and HeLa (Huang, Azgari, Yin, Chiou, Lindsey-Boltz, Sancar, Hu & Adebali, 2022) Damage-seq and XR-seq data were used to examine the damage formation tendencies and repair profiles on R-loops. *Arabidopsis* wild-type and mutant XR-seq data (Kaya, Adebali, Oztas & Sancar, 2022; Oztas, Selby, Sancar & Adebali, 2018) were used to analyze repair activity on *Arabidopsis* protein-coding genes and R-loops. Input DNA-seq data from NHF1 (SRA: SRR5461463) and HeLa (SRA: SRR11147234) cell lines were included when creating simulated Damage-seq and XR-seq data (Akkose & Adebali, 2023; Hu et al., 2017; Huang et al., 2022).

ATAC-seq data obtained from HeLa cell line was retrieved from Li et al. (Li, Meissner, Wang, Du, Ma, Kshirsagar, Tilburgs, Buenrostro, Uesugi & Strominger, 2021). Histone marker ChIP-seq data of HeLa cell line that was obtained by ENCODE project was retrieved from the NCBI website (PRJNA63443) (Consortium & others, 2012).

Simple somatic mutations in melanoma (MELA-AU), skin cutaneous melanoma (SKCM-US), skin adenocarcinoma (SKCA-BR), breast cancer (BRCA-US) and uterine corpus endometrial carcinoma (UCEC-US) genomes were re-

trieved from the International Cancer Genome Consortium (ICGC) data portal (https://dcc.icgc.org/releases/release_28).

ChIP-seq data of R-loop regulator proteins used in Hidden Markov Model (HMM) were listed in Table 3.1.

3.2 R-loop data processing

DRIP-seq, ssDRIP-seq, qDRIP-seq and RR-ChIP-seq raw data were retrieved and processed as described in their original articles (Crossley et al., 2020; Hamperl et al., 2017; Tan-Wong et al., 2019; Yang et al., 2019). From the final BED files, peak calling was performed using MACS2 algorithm (version 2.2.4) (Zhang, Liu, Meyer, Eeckhoute, Johnson, Bernstein, Nusbaum, Myers, Brown, Li & others, 2008). Peak centers were extended by 2 kb and 5 kb upstream and downstream for further analysis.

Level 7 R-loops were directly used from R-loopBase database (Lin et al., 2022). Since the positions were provided for ssDNA strands of the R-loops, the labelling was done accordingly.

Strand assignment of RLBase R-loops was done by intersecting them with the R-loopBase Level 5 R-loops from plus and minus strands. Strand information was given if the intersection between the two R-loop positions were more than 50 bases. R-loop centers were extended by 5 kb upstream and downstream for further analysis.

Spearman correlations and PCA plots were extracted using deepTools (version 3.5.1.).

3.3 Curation of G-quadruplex positions on human genome

To effectively map the genome-wide distribution of G-quadruplex (G4) structures, we employed a blend of experimental and computational techniques. Our approach included the use of G4P-ChIP-seq and G4-Cut&Tag methods to precisely locate G-quadruplex structures within the human genome, as cited in the studies by Li et

al. (Li, Wang, Yin, Fang, Xiao, Xiang, Wang, Li, Huang, Huang & others, 2021) and Zheng et al. (Zheng, Zhang, He, Gong, Wen, Chen, Hao, Zhao & Tan, 2020). In addition, we utilized computational tools such as G4-Miner and Quadron, as described by Sahakyan et al. (Sahakyan, Chambers, Marsico, Santner, Di Antonio & Balasubramanian, 2017) and Tu et al. (Tu, Duan, Liu, Lu, Zhou, Sun & Lu, 2021). These tools are specifically designed to detect G-quadruplex structures when applied to whole-genome sequencing data and the GRCh38 reference genome. For constructing a reliable G-quadruplex distribution map across the genome, we selected any G4 peak from the computational datasets that overlapped with at least one experimental peak.

In summary, for the G4-Cut&Tag method, we processed the raw data by aligning it to the GRCh38 reference genome using Bowtie2 (version 2.4.1) (Langmead & Salzberg, 2012), applying specific options such as "–no-unal –no-discordant –no-mixed –seed 1 –reorder". We excluded reads with a mapping quality below 20. Peak calling was then conducted using the MACS2 tool (version 2.2.7.1) (Zhang et al., 2008) with a set threshold of q = 0.0001. For the G4P-ChIP-seq process, we incorporated additional options "–sensitive-local –no-unal" to those used in G4-Cut&Tag. We also eliminated duplicate reads and those with low mapping quality (below 20) using samtools (version 1.10) rmdup (Li, Handsaker, Wysoker, Fennell, Ruan, Homer, Marth, Abecasis, Durbin & Subgroup, 2009). Finally, MACS2 was employed again, this time with the settings "-q 0.001 –keep-dup 1".

3.4 Histone ChIP-seq and ATAC-seq profiles on R-loops

Histone ChIP-seq and ATAC-seq data were processed as defined in their original studies (Consortiumet al. , 2012; Li et al., 2021). The processed reads were intersected with R-loops using BEDTools (version 2.30.0) intersect utility as well as deepTools (version 3.5.1) (Quinlan & Hall, 2010; Ramírez, Dündar, Diehl, Grüning & Manke, 2014).

3.5 R-loop distribution on chromatin states

Chromatin states for human genome were taken from ChromHMM study (Ernst & Kellis, 2017). Each segment in the states were intersected with the RR-ChIP-seq and qDRIP-seq data reads and intersecting reads were counted. After RPKM normalization, boxplots were generated with ggplot2 R package.

Arabidopsis chromatin states were retrieved from Sequeira-Mendes et al. (Sequeira-Mendes, Aragüez, Peiró, Mendez-Giraldez, Zhang, Jacobsen, Bastolla & Gutierrez, 2014). *Arabidopsis* ssDRIP-seq peaks were counted on each segment of the states and plotted using ggplot2 R package.

3.6 Transcription activity on R-loops

To assess the transcription on R-loops, GRO-seq data was used. Human GRO-seq data was retrieved from Andersson et al. (Andersson, Refsing Andersen, Valen, Core, Bornholdt, Boyd, Heick Jensen & Sandelin, 2014) and *Arabidopsis* GRO-seq data was obtained from Zhu et al. (Zhu, Liu, Liu & Dong, 2018). Both data were processed as described in the original articles. The profiles on human and *Arabidopsis* R-loops were plotted using deepTools (Ramírez et al., 2014).

3.7 Damage-seq and XR-seq data processing

NHF1 wild-type 0-hour, 1-hour, 8-hour and 24-hour Damage-seq and 1-hour, 8-hour and 24-hour XR-seq data, NHF1 CSB knock-out 1-hour XR-seq data, HeLa 0-hour Damage-seq and 12-minute XR-seq data, and *Arabidopsis* wild-type, CSA1 knock-out, CSA2 knock-out and double knock-out XR-seq data were processed using the codes available at https://github.com/CompGenomeLab/xr-ds-seq-snakemake. Briefly, Cutadapt (v4.1) (Martin, 2011) was employed to remove the adaptor sequence (TGGAATTCTCGGGTGCCAAGGAACTCCAGTNNNNNNACGATCTCGTATGCCGTCTTCTGCTTG) from the reads. The trimmed reads were aligned to the human (GRCh38) and *Arabidopsis* (Araport11) genomes using Bowtie2 (v2.4.1) (Langmead & Salzberg, 2012). The SAM files underwent quality filtering using samtools (v1.10) with a threshold of 20 (-q 20) and were subsequently

converted to BAM format (Li et al., 2009). Following the removal of duplicates using Picard (v2.27) (removes adapter sequences from high-throughput sequencing reads, 2019), BEDTools (v2.29.0) was employed to acquire BED files (Quinlan & Hall, 2010).

The Boquila algorithm (Akkose & Adebali, 2023) was utilized within the snakemake process to generate simulated Damage-seq and XR-seq data. Human data simulations were created with the input DNA-seq option.

3.8 Damage and repair profiles

To determine the damage and repair patterns, the R-loop site centers were expanded by 5 kilobases (kb) in both the upstream and downstream directions. The 10 kb sections were partitioned into 25-base pair (bp) segments using the BEDTools makewindows (Quinlan & Hall, 2010). The intersection of each bin was determined by checking and counting the plus and minus Damage-seq and XR-seq reads, as well as the simulated Damage-seq and XR-seq reads with the extended R-loop positions, using the BEDTools intersect function with the parameters -c and -F 0.51.

The number of overlaps on each bin was then subjected to RPKM (Reads Per Kilobase Million) normalization by adding a pseudocount of 1 to each count. The normalized repair was determined by dividing the RPKMs of Damage-seq read overlaps by the RPKMs of simulated Damage-seq read overlaps for each bin. Similarly, the RPKMs of XR-seq read overlaps were divided by the RPKMs of simulated XR-seq read overlaps on each bin to determine the normalized repair rates. The relative repair rates were calculated by dividing the normalized repair by the normalized damage. The profiles were graphed using the Python library seaborn (version 0.10.1) (Waskom, 2021).

3.8.1 Damage and relative repair on G4-containing and G4-lacking R-loops

G4 positions were intersected with RLBase R-loop ssDNA positions and R-loops that contained a G4 on at least 50% of its ssDNA were defined as 'R-loops with

G4s'. Using BEDTools subtract function (-A), R-loops that lack an overlapping G4 were defined as 'R-loops without G4s'.

The normalized damage calculations were performed on each R-loop as explained in Section 3.4. Boxplots were generated using the R package ggplot2 (Wickham, 2011).

3.8.2 Damage and relative repair on R-loop-containing and R-loop-lacking regions

The annotated protein-coding genes of the human genome (GRCh38) were obtained and any genes that overlapped with another gene was removed. The remaining set of genes was compared with the regions that include R-loop centers and 1 kilobase away from those centers in both directions. The segments of the genes that overlap with at least 80% of the R-loop center regions were designated as 'gene segments containing R-loops'. Genes lacking R-loops were chosen by employing the BEDTools subtract function (-A). The BEDTools shuffle tool was employed to randomly choose gene segments without R-loops from genes that were devoid of R-loops considering the number and the length of 'gene segments containing R-loops', to define a set of 'gene segments without R-loops'.

The normalized damage, normalized repair and relative repair rates were calculated on each gene segment as explained in Section 3.4. The boxplots were generated using the R tool ggplot2 (Wickham, 2011).

3.8.2.1 Cell culture

HeLa cells were cultured in Dulbecco's modified Eagle's medium (PAN-P04-03500), supplemented with 10% heat-inactivated fetal bovine serum (PAN-P30-3304), 100 U/ml penicillin/streptomycin (PAN-P06-07100), 2 mM L-glutamine (PAN-P04-80100), and 1x MEM non-essential amino acid solution (Gibco 11140-35). The cells were maintained at 37 °C with 5% CO2 until reaching 80% confluence.

3.8.2.2 RNA-seq library preparation & sequencing

RNA isolation was carried out using Genezol RNA isolation reagent (Geneaid GZR100) following the manufacturer's instructions. Library preparation and sequencing was performed by Novogene (UK) Co. Ltd. Briefly, quantification and qualification of the RNA, 1% agarose gels were used to monitor RNA degradation and contamination, while the NanoPhotometer® spectrophotometer (IMPLEN, CA, USA) was utilized to check RNA purity. Additionally, RNA integrity was assessed using the RNA Nano 6000 Assay Kit on the Bioanalyzer 2100 system (Agilent Technologies, CA, USA). For RNA sample preparations, 0.4 micrograms of RNA per sample was used. Sequencing libraries were generated using the NEBNext® UltraTM RNA Library Prep Kit for Illumina® (NEB, USA) as per the manufacturer's recommendations, with index codes added to attribute sequences to each sample. The process involved mRNA purification, fragmentation, first strand cDNA synthesis, and second strand cDNA synthesis. The remaining overhangs were converted into blunt ends, and NEBNext Adaptors were ligated. To select cDNA fragments of 250 300 bp in length, library fragments were purified using the AMPure XP system (Beckman Coulter, Beverly, USA). Subsequently, PCR was performed, and the PCR products were purified (AMPure XP system), with library quality assessed on the Agilent Bioanalyzer 2100 system. Index-coded samples were clustered on a cBot Cluster Generation System using TruSeq PE Cluster Kit v3-cBot-HS (Illumina) following the manufacturer's instructions. After cluster generation, the libraries underwent sequencing on an Illumina Novaseq 6000 platform, producing 150 bp paired-end reads.

3.8.2.3 RNA-seq data processing

After preprocessing and quality control of FASTQ files with fastp (Chen, Zhou, Chen & Gu, 2018), raw sequencing reads have been aligned to human genome (GRCh38) with STAR aligner version 2.7.6a (Dobin, Davis, Schlesinger, Drenkow, Zaleski, Jha, Batut, Chaisson & Gingeras, 2013) with arguments –genomeLoad NoSharedMemory –quantMode TranscriptomeSAM –twopassMode Basic –outSAMtype BAM SortedByCoordinate. Then, expression of transcripts (transcripts per million; TPM) was quantified using Salmon version 1.6.0 (Patro, Duggal, Love, Irizarry & Kingsford, 2017).

GRCh38 protein-coding genes were classified according to their expression levels.

Genes with TPMs 5 to 10 and with TPMs smaller than 1 were subjected to further analysis.

3.9 Mutation data processing

The somatic mutations of the five cancer cohorts (release 28) were obtained from the ICGC website (https://dcc.icgc.org/releases/release_28). The data underwent filtration, retaining only single base substitutions for subsequent study. The mutations were mapped to the GRCh38 genome using the UCSC liftOver tool (Kuhn, Haussler & Kent, 2013). The fasta files were acquired and the sequences were compared with the substitution attributions present in the files. Among the corresponding ones, substitutions from the nucleotide C to T were designated as mutations on the plus strand, while changes from the nucleotide G to A were designated as mutations on the minus strand.

An analysis was conducted on the distribution of mutations in genomic regions that include R-loop centers and the regions immediately adjacent to them, spanning a distance of 5 kb in both directions, in five different types of cancer. The 10 kb areas were partitioned into 25-base pair bins and overlapped with C>T mutations from each form of cancer. The number of overlaps was recorded and the C nucleotide composition of each bin was utilized to standardize the mutation counts, with the addition of 1 as a pseudocount. The profiles were generated using the seaborn python package Waskom (2021).

The C>T mutations were also compared with gene segments that either had or did not have R-loops as defined in Section 3.4.1., and the number of overlaps was recorded. The overlap counts were normalized using the C content, with an added pseudocount of 1. The boxplots were generated using the R tool ggplot2.

The observed and expected C>T mutations were calculated and graphed using the codes supplied in Frigola et al. (Frigola, Sabarinathan, Mularoni, Muiños, Gonzalez-Perez & López-Bigas, 2017) on the identical set of gene segments with or without R-loops.

3.10 Molecular dynamics (MD) simulations

The crystal structure of DNA-RNA duplex (1G4Q) with the DNA sequence CTTTTCTTTG was downloaded from the Protein Data Bank (www.rcsb.org) (Berman, Westbrook, Feng, Gilliland, Bhat, Weissig, Shindyalov & Bourne, 2000). It was cleaned by removing the crystal waters and checked for any abnormalities such as any extra heterogeneous atom or missing base in Pymol (www.pymol.org) (DeLano & others, 2002). The complex was separated to obtain the single stranded DNA (ssDNA) chain. A double stranded DNA-DNA complex was prepared using the ssDNA as template using Biovia Discovery Studio (Systèmes, 2020).

All complexes were subjected to classical molecular dynamics studies using GROMACS (2020) (Abraham, Murtola, Schulz, Páll, Smith, Hess & Lindahl, 2015; Van Der Spoel, Lindahl, Hess, Groenhof, Mark & Berendsen, 2005a), on TESLA K-80 graphical processor unit (GPU) installed on the TOSUN high performance computing (HPC) server at Sabanci University.

The complexes were prepared using the AMBER99sb forcefield (Cheatham III & Case, 2013; Showalter & Brüschweiler, 2007). TIP3 water model (Jorgensen, Chandrasekhar, Madura, Impey & Klein, 1983) was used to enclose the complexes in a water box and ions were added to neutralize the system. The molecules were minimized for 100 picoseconds (ps) and then, subjected to equilibration at constant volume and temperature of 310°K (NVT) and 1 mm Hg pressure (NPT) for 200 ps, using V-rescaling and Parrinello-Rahman methods, respectively (Bussi, Donadio & Parrinello, 2007; Parrinello & Rahman, 1981; Rühle, 2008). 500-nanosecond (ns) simulations were done for each molecule in triplicate and the coordinates were recorded at every 20.0 ps. Leapfrog integrator (Berendsen, 1986) was used for capturing the equation of motion. Particle-mesh Ewald (PME) (Darden, York & Pedersen, 1993) conditions were applied to calculate the long-range interactions with a cutoff of 1 nm. Short term Coulombic and van der Waals interactions cutoff was kept at 1 nm.

Additional position restraints were added at the guanine base at the end of the ssDNA structure while producing the simulation to prevent it from forming hairpin structures or interacting with the cytosine (DG1) at the 5' end. A separate index file and a position restraints file were created for the guanine (DG10) atoms using GROMACS built-in functions (Van Der Spoel, Lindahl, Hess, Groenhof, Mark & Berendsen, 2005b).

GROMACS built-in analysis tools were used to calculate the root mean squared deviation (RMSD) over time and base-wise root mean squared fluctuations (RMSF). VMD (Humphrey, Dalke & Schulten, 1996) was used to calculate the pair-wise atomic distances and the dihedral angles between C5 and C6 of two adjacent thymine

(T) bases (DT2 and DT3) in all the complexes over the course of the trajectory. VMD was also used to calculate the residue-residue pair and unique hydrogen bond formation. Normal mode analysis was done using the algorithm developed by Midstlab available on Github (https://github.com/midstlab/md_simulation_analysis) for all the complexes. VMD was used to visualize the major movements of the molecules resulting from normal mode analysis. Chimera (2023) (Chimera, 2004) visualization software was used for observing the final structures and obtaining the figures.

3.11 HMM prediction of genomic states

The available human ChIP-seq data of R-loop regulator proteins were searched from the literature and retrieved without respect to cell type (Table 3.1). Only the the ChIP-seq experiments performed in wild-type cell lines with no treatment were selected. Control datasets were also included in the analysis if available. Raw ChIP-seq data were trimmed using fastp (Chen et al., 2018) and aligned on human genome GRCh38 using Bowtie2 (Langmead & Salzberg, 2012). The resulting SAM files were converted to BAM format by quality-trimming (-q 20) with samtools (Li et al., 2009) and duplicate reads were removed using Picard toolkit (Pic, 2019). Convertion from BAM to BED format and sorting were performed with BEDTools (Quinlan & Hall, 2010).

For the genome state analysis, ChromHMM tool was used (Ernst & Kellis, 2017). The GRCh38 assembly was divided into 5kb-sized bins, and ChromHMM was employed to identify peaks in samples associated with the selected regulators. Peaks were marked as 1, while non-peaks were designated as 0. Subsequently, all bins were indexed with the chosen regulators, and the ChromHMM Markov chain prediction model, employing default settings, was applied to identify the optimal distribution of chromatin states, with the selection of number of states as 10.

Each bin in the states were classified according to their R-loop content. Using BEDTools intersect command with option (-F 0.8), we have overlapped each 5 kb-sized bin with R-loop centers and -/+ 1 kb flanking regions. The bins that include at least 80% of the R-loop regions were defined as 'bins with R-loops'. The bins that had no overlap with the whole R-loop peaks (BEDTools subtract -A option) were classified as 'bins without R-loops'. ATAC-seq (Li et al., 2021) read abundances

were calculated on each bin and subjected to RPKM normalization. Relative repair was also calculated as described earlier. Normalized ATAC-seq and relative repair was calculated by diving the ATAC-seq RPKM and relative repair on each bin with R-loops by the mean of all bins without R-loops within each state.

Table 3.1 ChIP-seq data used in HMM predictions

Regulator protein	Reference
ATR	Li et al., 2018; Solvie et al., 2022
BRCA1	Heidari et al., 2014; ENCODE project
CTCF	Ibarra et al., 2016
DDX21	Johansson et al., 2020; Calo et al., 2018
FIP1L1	Heidari et al., 2014
FUS	Xiao et al., 2019
NFAT5	Heidari et al., 2014
PRMT5	Moyers et al., 2023
RAD51	Dellino et al., 2019
RPA1	Zheng et al., 2020
SETX	Miller et al., 2015
SRSF1	Xiao et al., 2019; Heidari et al., 2014
SRPK2	Sridhara et al., 2017
XRN2	Brannan et al., 2012
SMC3	Spracklin et al., 2021
SIRT7	Vazquez et al., 2019
U2AF1	Xiao et al., 2019
WRN	- (GSE68714)

3.12 Evolutionary analysis of *Arabidopsis* CSA proteins

The BLAST algorithm (version 2.9.0) (Camacho, Coulouris, Avagyan, Ma, Papadopoulos, Bealer & Madden, 2009) was employed to conduct a comparison between all eukaryotic protein sequences in the Uniprot database and CSA1 and CSA2, individually. The comparison was performed with an e-value threshold of 1E-06 and a maximum of 1000 target sequences. The protein hits for CSA1 and CSA2 were consolidated and duplicate sequences were removed. The MAFFT (version 7.407) (Katoh & Standley, 2013) software was utilized to accomplish multiple sequence alignment, employing the –auto option. The conservation scores for the residues in the multiple sequence alignment were computed using Bio3D (Grant, Rodrigues,

ElSawy, McCammon & Caves, 2006). A phylogenetic tree was constructed using IQ-TREE2 (version 2.0.6) (Nguyen, Schmidt, Von Haeseler & Minh, 2015) with the 200 most comparable eukaryotic protein sequences to CSA1 and CSA2. The "-B 1000" option was utilized to acquire bootstrap values.

4. RESULTS

4.1 Distribution of R-loops on human genome

4.1.1 Comparison of R-loop sequencing methods and assessment of data processing pipeline

R-loops are dynamic structures on genomes that mostly form during transcription (Hegazy et al., 2020). Even though they are prevalently found on genic regions, intergenic R-loops also exist (Graf, Bonetti, Lockhart, Serhal, Kellner, Maicher, Jolivet, Teixeira & Luke, 2017). Therefore, we aimed to assess the genomic distributions of R-loop sites in order to understand more about R-loops and the data in our hands. Firstly, we have retrieved DRIP-seq (Hamperl et al., 2017) and strand-specific ssDRIP-seq (Yang et al., 2019), qDRIP-seq (Crossley et al., 2020) and RR-ChIP-seq (Tan-Wong et al., 2019) data that were obtained from HeLa cells and aligned them on human genome GRCh37 (Figure 4.1). DRIP-seq and ssDRIP-seq data showed a relatively homogeneous distribution on the chromosomes while there were higher peaks on some regions with qDRIP-seq and RR-ChIP-seq data (Figure 4.1). On the other hand, the peaks of qDRIP-seq and RR-ChIP-seq data were not consistent with each other, emphasizing the differences between the four methods that needed to be further analyzed.

Figure 4.1 Distribution of DRIP-seq, ssDRIP-seq, qDRIP-seq and RR-ChIP-seq reads on human genome (GRCh37) chromosome 1. Processed reads were aligned on 5 kb bins of chromosome 1 and intersections were subjected to RPKM normalization in each bin.

To compare the methods and at the same time, to assess our data processing pipeline, we have compared the processed data provided by the studies with the processed data with our pipeline as well as the data from different methods. For that reason, we have performed PCA analyzes and calculated Spearman correlations to compare the genome coverage patterns between the data. In the PCA analysis, RR-ChIP-seq data processed by the original study and by us were located quite close to each other while ssDRIP-seq data processed by the original study was located relatively distant to the data processed by us (Figure 4.2). In addition, while ssDRIP-seq and qDRIP-seq methods were clustered together, RR-ChIP-seq data was located further away from other data.

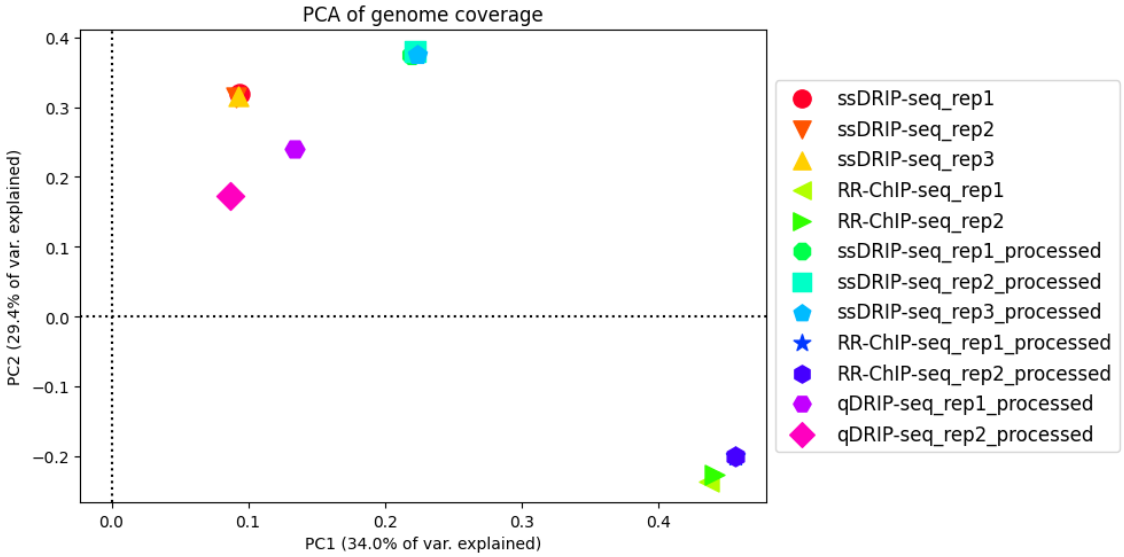

Figure 4.2 PCA plot for the genome coverage by different methods and processing pipelines. Data labelled as 'processed' refers to the data processed by our in-house pipeline.

We have also computed Spearman correlations between the data from different methods. ssDRIP-seq data replicates processed by us and the original study were clustered together with high correlation values whereas qDRIP-seq and RR-ChIP-seq were more similar to each other (Figure 4.3).

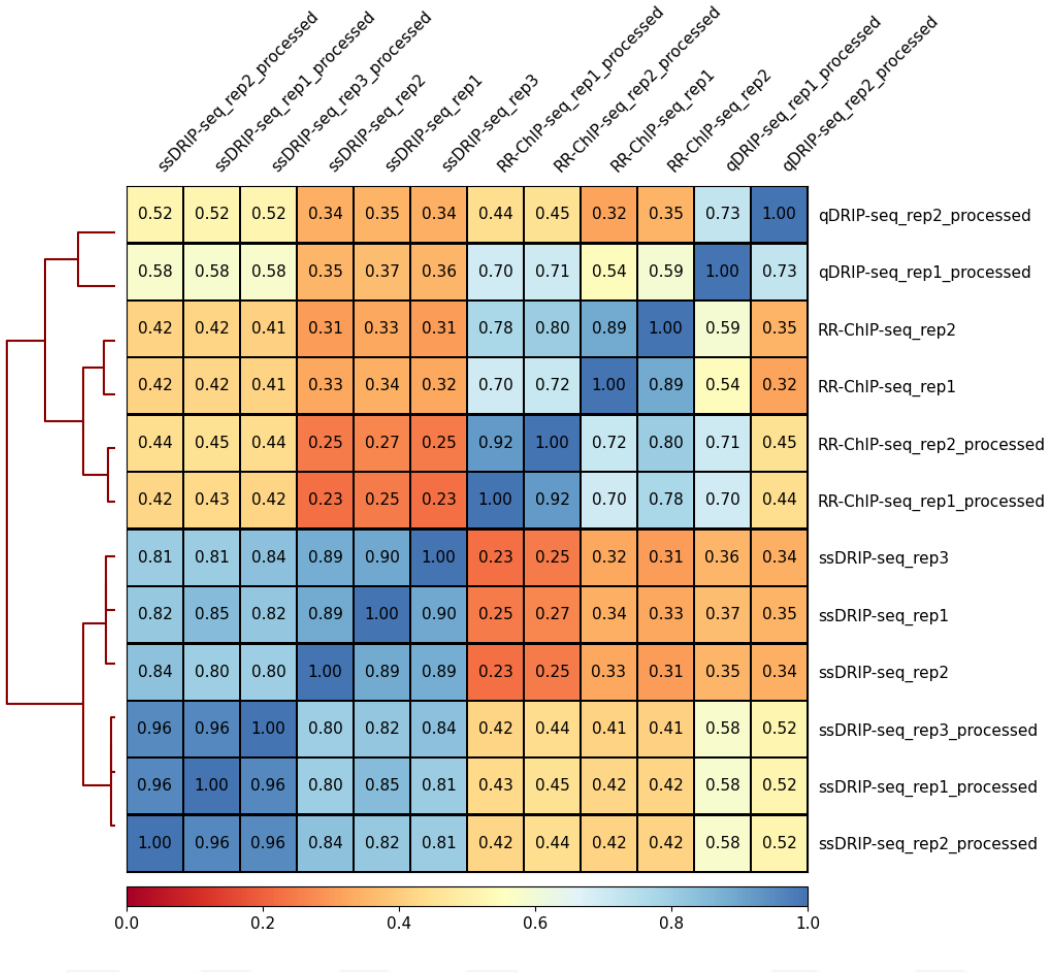

Figure 4.3 Heatmap of Spearman correlations of genome coverage between data from different methods and processing pipelines. Data labelled as 'processed' refers to the data processed by our in-house pipeline.

Since the provided processed reads of qDRIP-seq study was aligned on human genome GRCh38 while our analysis was on GRCh37, the data provided by the original article was not included in the PCA and correlation analyzes (Figure 4.2, 4.3). Instead, we have processed the qDRIP-seq raw data by aligning the reads on GRCh38 and compared the genome coverage separately (Figure 4.4). The similarity was not great but sufficient between our pipeline and the processing of the original study. Altogether, these results suggested that although all of these methods aimed to capture R-loop positions from HeLa cell line, the resemblance between the data was very low which could be due to the difference in antibody-based and RNase H-based methods, the specificity of S9.6 antibody, quality issues of the data or the nature of R-loops being formed and resolved continuously. Due to the genomic distribution of the reads with higher intensity peaks instead of a more homogeneous distribution, we decided to continue with qDRIP-seq and RR-ChIP-seq data at this point. On the other hand, we were convinced that our data processing was well

enough to continue with the further analyzes.

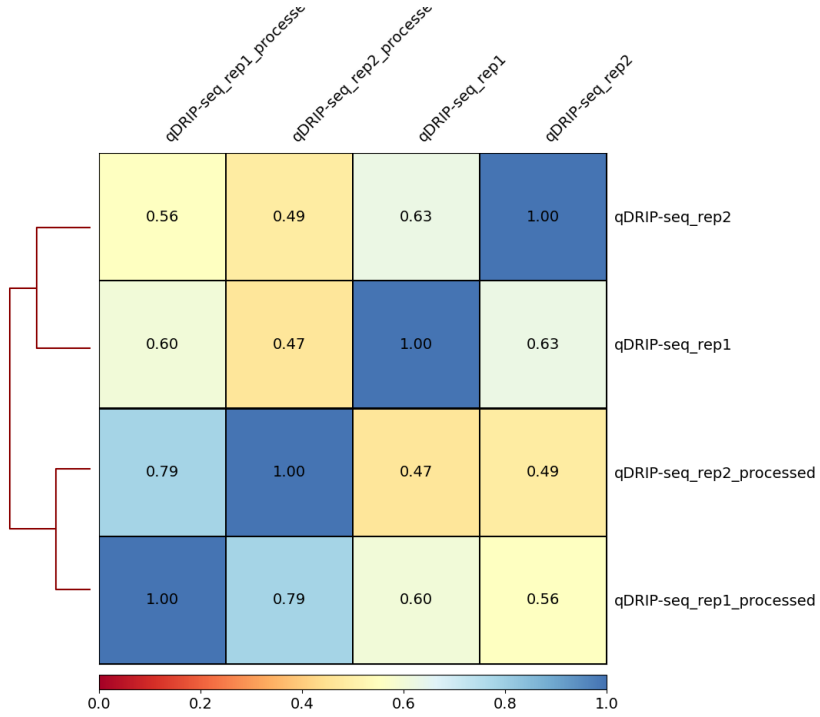

Figure 4.4 Heatmap of Spearman correlations of genome coverage of qDRIP-seq data processed by the original article and by our pipelines. Data labelled as 'processed' refers to the data processed by our in-house pipeline.

4.1.2 R-loops on genic regions

R-loops are mostly formed around genic regions since they are the products of transcription (Hegazy et al., 2020). In order to assess the data in terms of distribution around genes, we have intersected regions around transcription start sites (TSS) and transcription end sites (TES) with the qDRIP-seq reads (Figure 4.5). qDRIP-seq reads, which determine the locations where DNA:RNA hybrid of the R-loops are formed, was peaking on TS at the downstream regions of TSS and upstream of TES as expected from the common way of R-loop formation where the nascent RNA anneals on the TS making the DNA:RNA hybrid. Read abundance was higher on NTS upstream of the TSS which could indicate antisense transcription on those regions.

Figure 4.5 Distribution of qDRIP-seq reads around TSS and TES. -/+ 2 kb regions around TSS (left panel) and TES (right panel) were intersected with qDRIP-seq reads strand-specifically.

Since R-loops are products of transcription, transcription level should be correlated with R-loop abundance. To test this, we have divided protein-coding genes into five classes according to their transcription levels. The first class included genes with no transcription while other four classes included equal number genes with increasing transcription levels. Genes with no expression did not have any difference between qDRIP-seq read abundance on TS and NTS (Figure 4.6). Similarly, '0-25%' genes which was the subgroup with the lowest expression did not show any read abundance difference on TS and NTS. On the other hand, '75-100%' genes with the highest expression contained the highest read abundance on TS of both downstream of the TSS and upstream of the TES. The same results were obtained using RR-ChIP-seq reads. These results confirmed that the expression and R-loop formation is correlated.

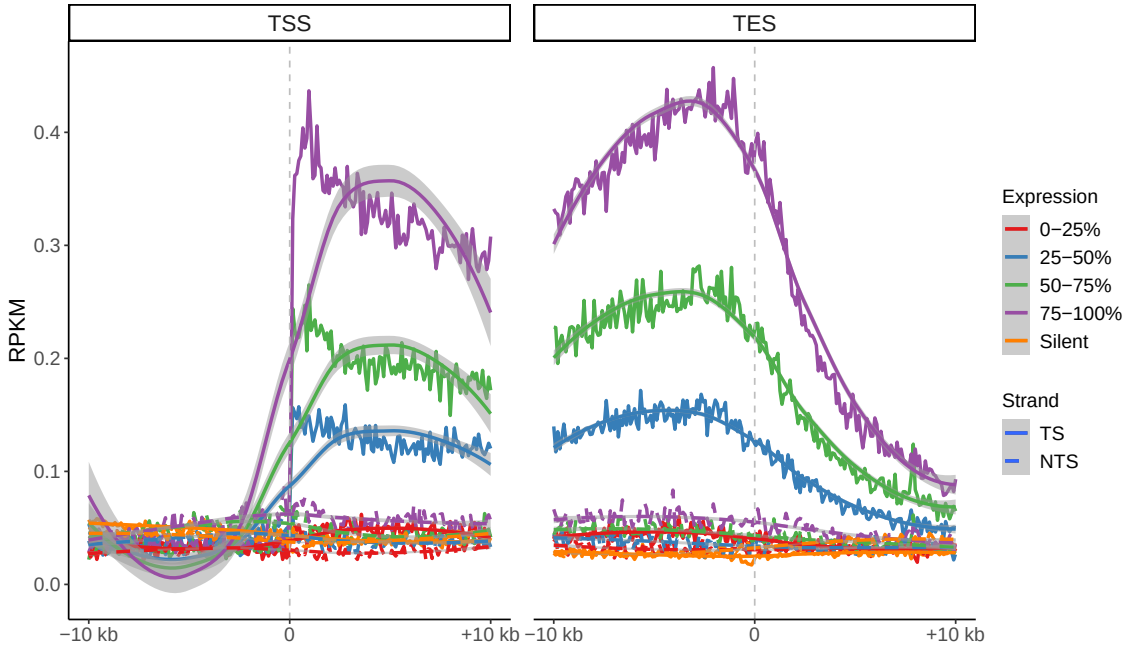

Figure 4.6 Distribution of qDRIP-seq reads around TSS and TES of genes with different expression levels. Protein-coding genes were subgrouped according to expression levels and -/+ 10 kb regions around TSS (left panel) and TES (right panel) were intersected with qDRIP-seq reads strand-specifically. Genes with no expression was termed as 'Silent. Percentages indicate the four subgroups of expression.

To assess R-loop formation on other genomic regions, we have used ChromHMM (Ernst & Kellis, 2017) regions and overlapped them with the qDRIP-seq and RR-ChIP-seq reads (Figure 4.7). Even though the levels of read abundances differed, the general trend of the distributions among ChromHMM states were alike. The highest read abundances were accumulated on genic regions, TSS regions and promoters in both of the data.

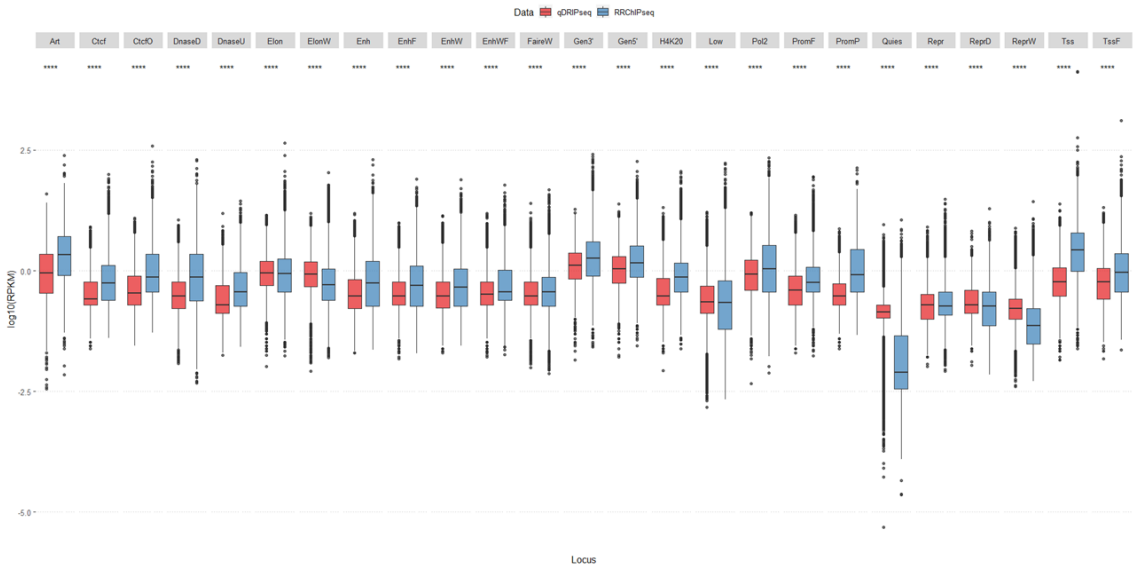

Figure 4.7 Distribution of R-loops on diverse chromatin structures. ChromHMM regions were intersected with qDRIP-seq and RR-ChIP-seq reads. Overlapped read counts were subjected to RPKM normalization. t-test was used to compute P-values (ns: p > 0.05; *: p <= 0.05; **: p <= 0.01; ***: p <= 0.001; ****: p <= 0.0001)

Since R-loop formation is correlated with the expression level, we expected R-loops to be mostly on open chromatin regions. To test this, we have retrieved and processed an ATAC-seq data obtained from HeLa cells and checked the ATAC-seq read distribution on qDRIP-seq and RR-ChIP-seq peaks (Figure 4.8). ATAC-seq read levels peaked within the 0.5 kb upstream of the RR-ChIP-seq R-loops while no peaks were observed on qDRIP-seq R-loops.

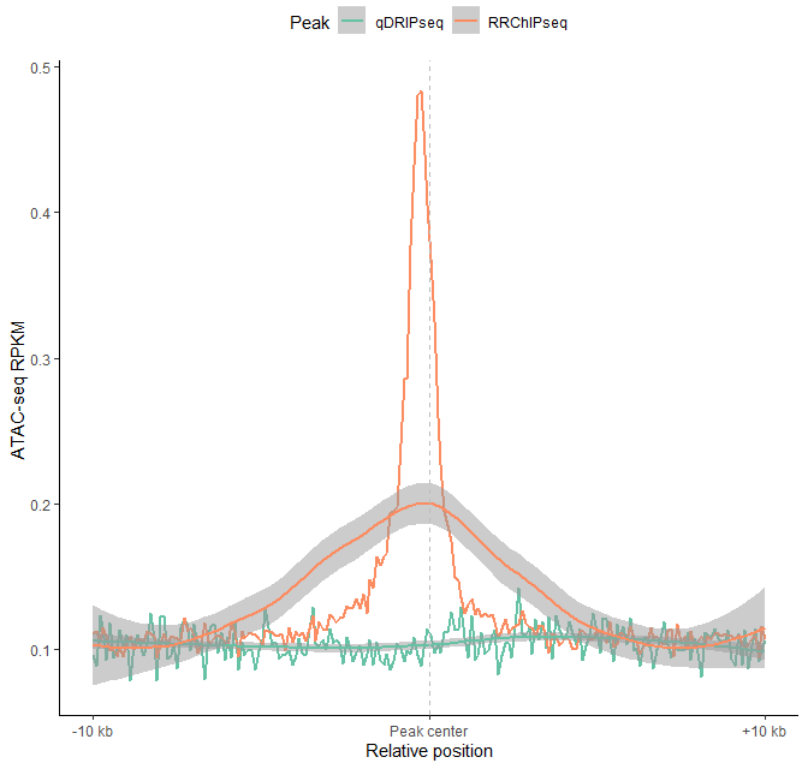

Figure 4.8 Distribution of ATAC-seq reads on qDRIP-seq and RR-ChIP-seq R-loops. ATAC-seq reads were intersected with the genomic bins on R-loop peak centers and -/+ 10 kb surrounding regions. Intersected read counts were subjected to RPKM normalization.

Histone markers give idea about various genomic features such as transcription activity or accessibility of the chromatin. Therefore, we checked the distributions of eight histone markers on qDRIP-seq and RR-ChIP-seq R-loops (Figure 4.10, 4.9). Among the eight histone markers, H3K4me3, H3K9ac and H3K27ac were the highest on RR-ChIP-seq peak centers while on qDRIP-seq peak centers, H3K36me3 and H3K79me2 were the highest (Figure 4.10, 4.9). In general, no similarities were observed between the peaks of RR-ChIP-seq and qDRIP-seq data in terms of histone marker distribution.

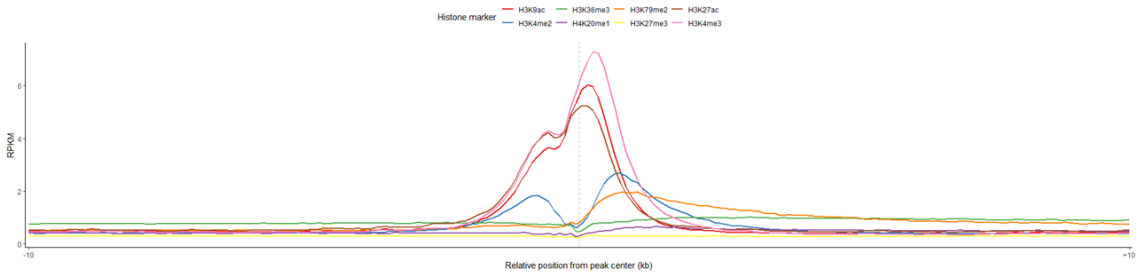

Figure 4.9 Distribution of histone markers on RR-ChIP-seq R-loops. Peak centers and -/+ 10 kb surrounding regions were included in the analysis. ChIP-seq reads of histone markers were overlapped with the R-loop peaks and intersected read counts were subjected to RPKM normalization.

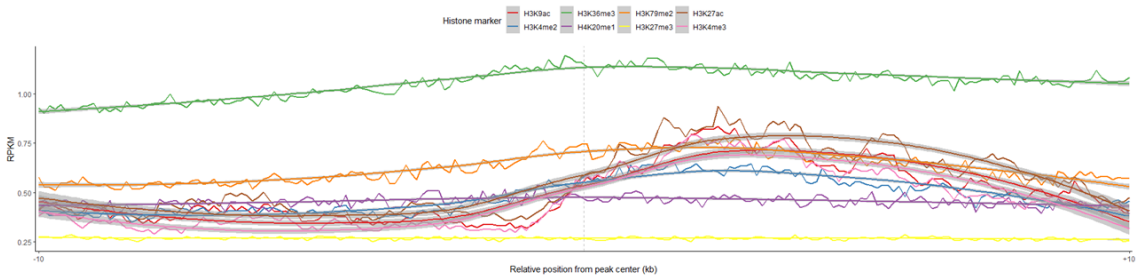

Figure 4.10 Distribution of histone markers on qDRIP-seq R-loops. Peak centers and -/+ 10 kb surrounding regions were included in the analysis. ChIP-seq reads of histone markers were overlapped with the R-loop peaks and intersected read counts were subjected to RPKM normalization.

Assessing R-loops from ssDRIP-seq, qDRIP-seq and RR-ChIP-seq methods revealed important differences among them. Even though their distribution on ChromHMM states was alike, they did not correlate in genomic distribution and coverage, accessibility, and histone marker occupancy. Thus, due to these inconsistencies, we decided to use neither of these data. Instead, we continued with the R-loopBase database(Lin et al., 2022) that was published recently at that time.

R-loopBase database (Lin et al., 2022) has gathered all of the published human R-loop datasets and compared the presence of each R-loop location within these datasets. The database provided nine levels of R-loops where level 1 included R-loops that were found in at least one dataset and level 9 included R-loops that were found in at least nine datasets. First, since level 9 corresponded to the highest confidence level, we have chosen level 9 to proceed with. However, the number of R-loops in level 9 was not convenient for further analysis (200 R-loops); therefore, we have chosen level 7 and level 8 which contained more R-loops.

In order to have an idea about R-loopBase R-loop levels, we have intersected them with human protein-coding gene locations. We have seen that in both level 7 and

level 8, 36% of R-loops were intersecting with a gene while 63% of them were intergenic. The genes intersecting with R-loops were mostly located at the ssDNA strand of the R-loops which is expected since R-loopBase database provides the locations of R-loop ssDNAs which are mostly located on the NTS of genes. When we looked at the transcription start sites (TSS) of these genes, we saw that most of the TSS were located at the upstream of R-loop peak center. At least one gene was located at both upstream and downstream regions of most of the ssDNAs while no genes were located at the sides of most of the DNA:RNA hybrid strands of level 7 R-loops (Figure 4.11). This result was expected since R-loop DNA:RNA hybrid strand generally forms when the nascent RNA anneals on the TS of the genes, leaving the NTS as single-stranded DNA (Costantino & Koshland, 2015; Hegazy et al., 2020).

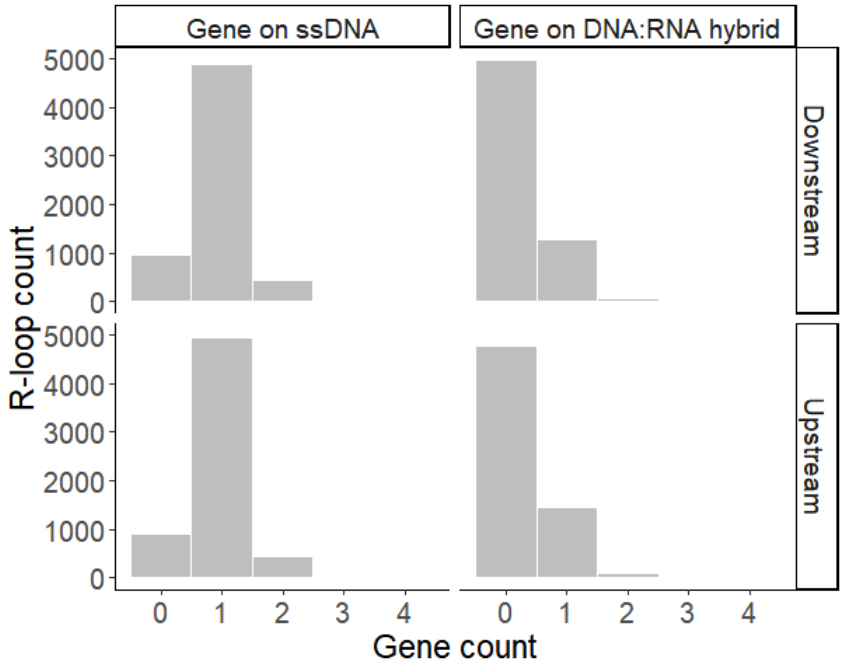

Figure 4.11 Genes counts intersecting with the upstream and downstream of R-loopBase R-loop strands. 2 kb upstream and downstream regions around R-loop peaks were intersected with protein-coding genes and the overlaps were counted.

The mean lengths of level 7 and level 8 R-loops were 391 and 315 nucleotides, respectively. The nucleotide content of levels was also checked. As expected, GC content increased as the level number increased. Since R-loop formation is correlated with the expression level, we expected R-loops to be mostly on open chromatin regions. To assess the R-loopBase R-loops in terms of chromatin accessibility, ATAC-seq read distribution on level 7 R-loops was checked (Figure 4.12). ATAC-seq reads were peaked within around 0.5 kb upstream region of the R-loop centers while R-loop peaks were also very high in read abundance. The distribution on R-loopBase R-loops was consistent with the distribution on RR-ChIP-seq R-loops, indicating

that these sets of R-loops are located in open chromatin regions, while R-loops captured by qDRIP-seq might have different properties.

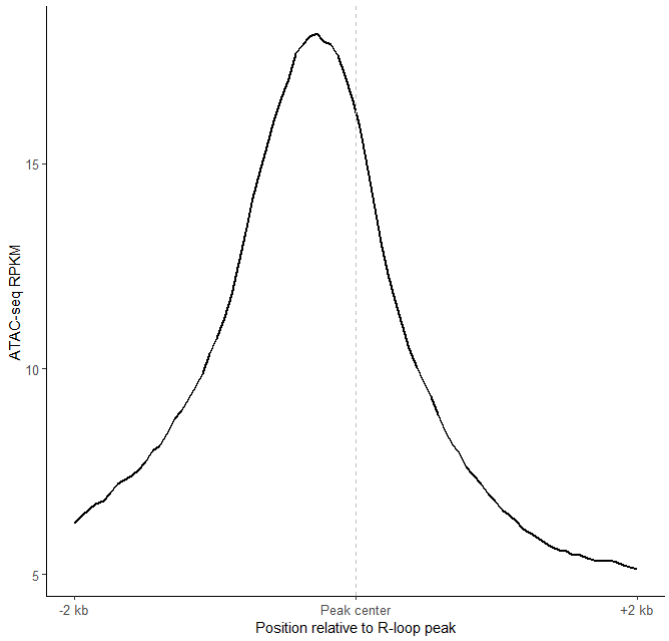

Figure 4.12 Distribution of ATAC-seq reads on R-loopBase R-loops. ATAC-seq reads were intersected with the genomic bins on level 7 R-loop peak centers and -/+ 2 kb surrounding regions. Intersected read counts were subjected to RPKM normalization.

After R-loopBase, RLBase database has been published which included human R-loop sites that were curated by gathering all human R-loop studies from various cell lines and assessing those R-loop sites using a set of quality-control criteria (Miller et al., 2022). Then, high-quality R-loop data were compared with each other to extract the 'consensus R-loop regions' (RLBase R-loops, hereafter). To check the distribution of RLBase R-loops on human genome GRCh38, we have first assessed the abundance of four histone markers: (1) H3K4me3, marking transcriptional start sites (TSS) (Benayoun, Pollina, Ucar, Mahmoudi, Karra, Wong, Devarajan, Daugherty, Kundaje, Mancini & others, 2014); (2) H3K36me3, marking actively transcribed regions and double-strand break (DSB) repair (Sun, Zhang, Jia, Fang, Tang, Wu & Fang, 2020); (3) H3K9me3 and (4) H3K27me3, marking heterochromatin and inhibited transcription (Becker, Nicetto & Zaret, 2016; Liu, Ali & Zhou, 2020) (Figure 4.13). The open chromatin marker H3K4me3 peaked starting from the R-loop start site until the end of the R-loops. The other open chromatin marker, H3K36me3, was the highest at the R-loop start site. The H3K36me3 abundance was lower on the R-loop than its upstream regions. Both heterochromatin markers, H3K9me3 and H3K27me3, were high on the R-loop start and end sites whereas being decreased towards R-loop centers.

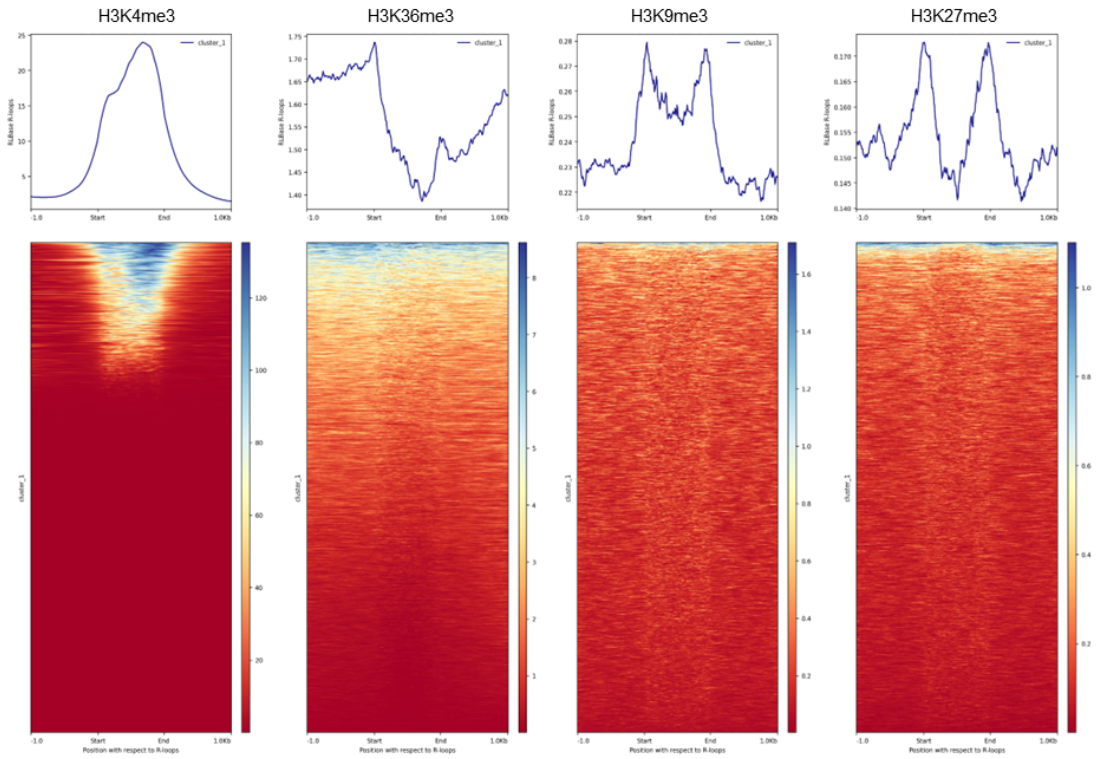

Figure 4.13 Abundance of histone markers on RLBase R-loops. ChIP-seq data of H3K4me3, H3K36me3, H3K9me3 and H3K27me3 markers were aligned on R-loops and 1 kb flanking regions upstream and downstream using deepTools (Ramírez et al., 2014). RPKM normalization was performed on overlapping read counts. 'Start' and 'End' sites on the plots refer to the start and end of the R-loop peaks.

RLBase R-loops were also rich in ATAC-seq reads (Figure 4.14a). The ATAC-seq reads were enriched at the R-loop bodies rather than the flanking 1 kb regions. However, when the R-loops were divided into two clusters in terms of read distribution, only around 20% of the R-loops showed a high read abundance difference between R-loop body and surrounding regions. The ATAC-seq read abundance did not peaked on R-loop bodies as high as the first cluster although a smaller peak was observable (Figure 4.14b). These results indicated that not all of the R-loops are same in terms of chromatin accessibility. Some R-loops might be located at more open chromatin than others.

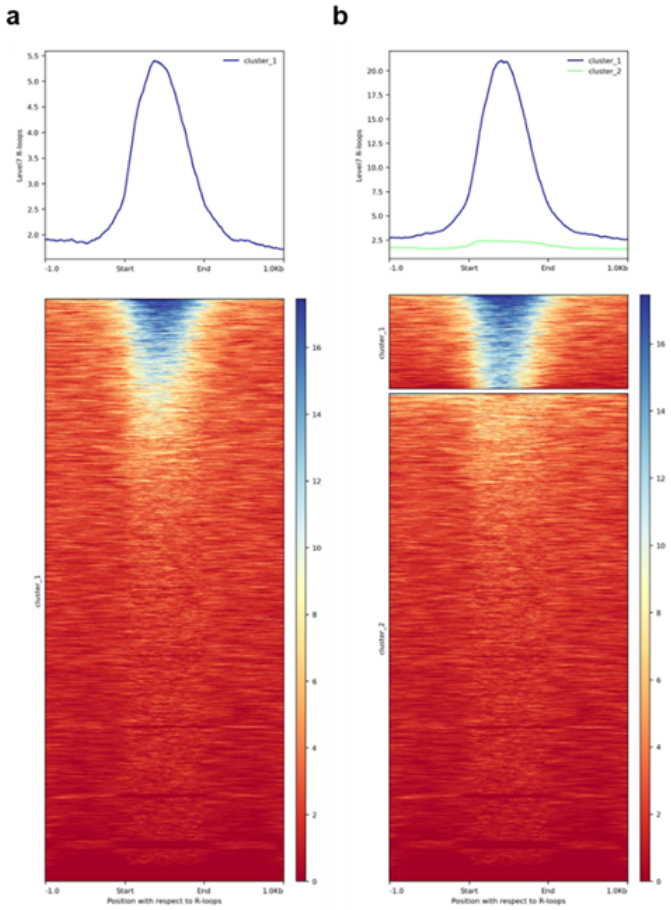

Figure 4.14 Abundance of ATAC-seq reads on RLBase R-loops. (a) ATAC-seq reads were aligned on R-loops and 1 kb flanking regions upstream and downstream using deepTools (Ramírez et al., 2014). RPKM normalization was performed on overlapping read counts. 'Start' and 'End' sites on the plots refer to the start and end of the R-loop peaks. (b) ATAC-seq reads were aligned on R-loops as explained in (a). Plotting was done with the option -kmeans 2 to differentiate R-loops at different clusters for read distribution.

RLBase R-loops were also assessed for their overlap with TSS regions (Figure 4.15). A high TSS content was observed on the start of R-loop ssDNA strands which decreased toward the end of the R-loops (Figure 4.15a). On the other hand, on DNA:RNA hybrid strand, no such abundance was observed (Figure 4.15b), confirming that most of RLBase R-loops were cis R-loops formed around the regions where the transcription occurred. In addition, TSS abundance on ssDNA confirmed that most of the RLBase R-loops were formed through the common way of R-loop formation in which ssDNA forms at the NTS.

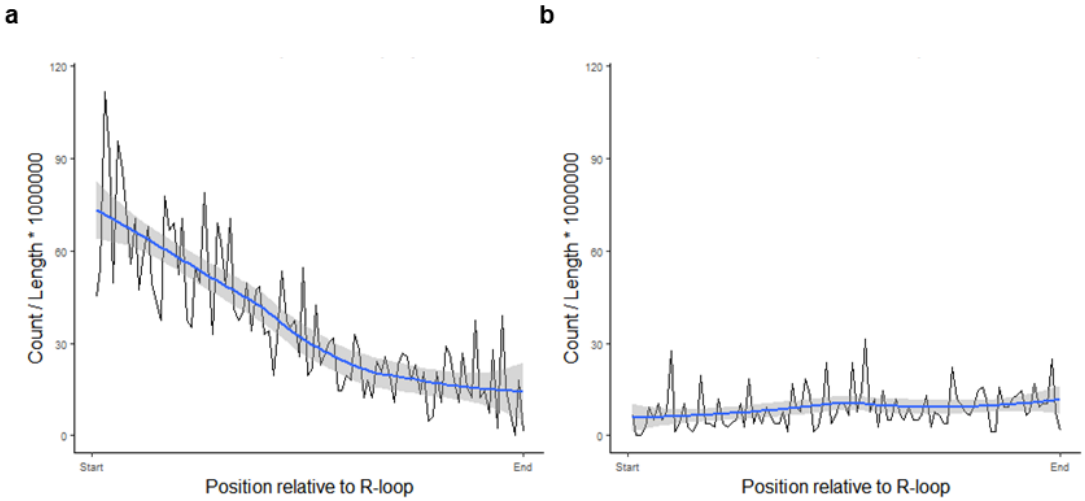

Figure 4.15 Abundance of transcription start sites on RLBase R-loops. The abundance of the first nucleotides of protein-coding genes on R-loop (a) ssDNAs and (b) DNA:RNA hybrids. 'Start' and 'End' sites on the plots refer to the start and end of the R-loop peaks.

Distribution of GRO-seq reads were also consistent with the distribution of TSS on RLBase R-loops (Figure 4.15, 4.16). GRO-seq reads on ssDNA strands were almost 2-folds higher than the read abundance on DNA:RNA hybrid strands (Figure 4.16a, b). GRO-seq method captures and sequences the nascent RNAs which should align on the NTS during genome alignment. Since most of the R-loop ssDNAs form on NTS of the genes, the higher abundance of GRO-seq reads on ssDNA coincided with the canonical formation of R-loops during transcription.

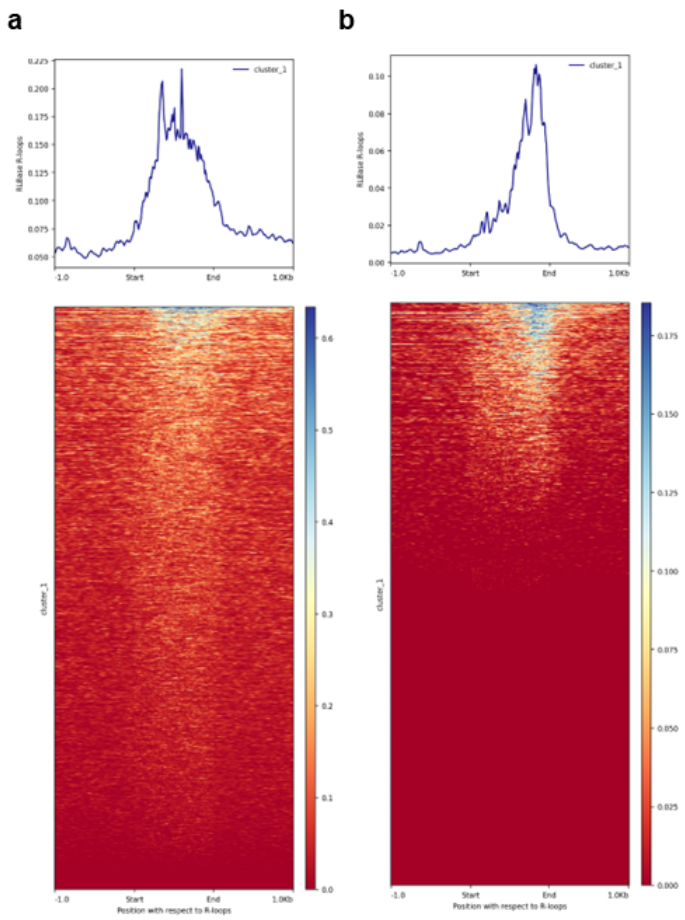

Figure 4.16 Abundance of GRO-seq reads on RLBase R-loops. GRO-seq reads were aligned on R-loops and 1 kb flanking regions upstream and downstream using deepTools (Ramírez et al., 2014). RPKM normalization was performed on overlapping read counts. 'Start' and 'End' sites on the plots refer to the start and end of the R-loop peaks. Distribution of GRO-seq reads on (a) ssDNA and (b) DNA:RNA hybrid strand was plotted separately.

Finally, we checked the length distribution of RLBase R-loops around their centers. The profile revealed that most of the RLBase R-loops covered -/+ 2 kb region spanning their centers (Figure 4.17). The mean of the lengths of the RLBase R-loops were 2523.35 while the median length was 1960 nucleotides.

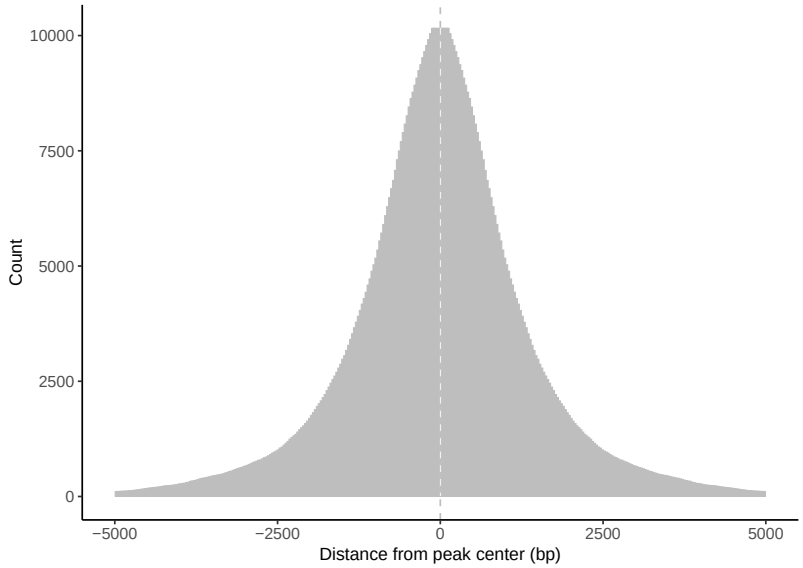

Figure 4.17 Abundance of RLBase R-loops around R-loop centers.

As the RLBase database provided R-loops that successfully passed a set of quality-control criteria and the initial results for their genomic distribution were as expected, we decided to continue with the RLBase R-loops for further analysis.

4.2 Damage and repair profiles on human R-loops

4.2.1 Damage profiles on R-loops

Ultraviolet (UV) light causes significant DNA damage by triggering the creation of cyclobutane pyrimidine dimers (CPDs) and 6-4 pyrimidine-pyrimidone photoproducts ((6-4)PPs). If left unrepaired, these DNA lesions hinder the normal process of DNA replication and significantly contribute to the occurrence of mutations in skin malignancies (Mao, Smerdon, Roberts & Wyrick, 2016). Therefore, it is important to know how UV-induced lesions were distributed on our genomes to identify the regions that are more prone to or more protected from the formation of CPDs and (6-4)PPs. Damage-seq and HS-Damage-seq methods were developed for mapping CPD and (6-4)PP lesions on the genome strand specifically with high sensitivity (Hu et al., 2017). We have mapped the UV-induced lesions on the genome, specifically

at sites where R-loops were detected as the damage formation tendency on R-loops has never been established clearly before.

As being the strand-specific data that we started to use in the beginning, we first checked the CPD and (6-4)PP damage distribution on ssDRIP-seq (Yang et al., 2019) peaks using the Damage-seq data obtained from the HeLa cells (Huang et al., 2022). We observed a higher damage abundance on the DNA:RNA hybrid strand and lower damage abundance on the ssDNA strand than their flanking regions for both damage types (Figure 4.18a, b). The levels of damage abundance were similar for both CPD (Figure 4.18a) and (6-4)PP lesions (Figure 4.18b).

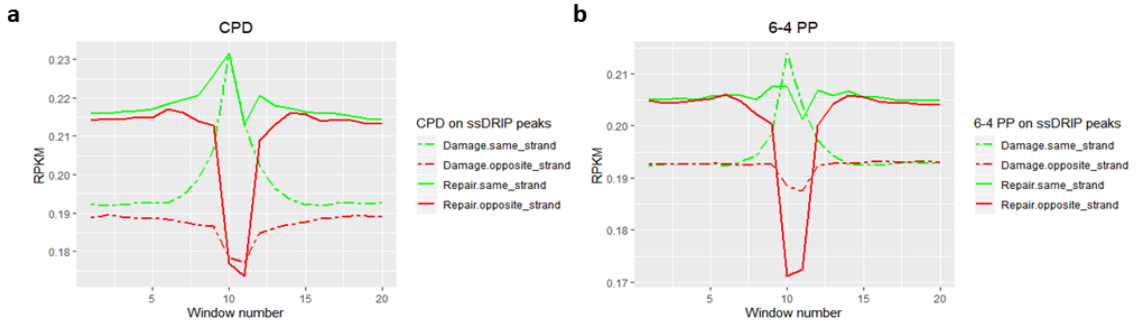

Figure 4.18 Damage-seq and XR-seq read distributions on ssDRIP-seq R-loops in HeLa cells. Damage-seq (0-hour time-point) and XR-seq (12-minute time-point) data in HeLa cells were used to extract read distributions on ssDRIP-seq R-loops centers and -/+ 10 kb flanking sites. 20 kb regions were divided into 20 bins and Damage-seq and XR-seq read counts were subjected to RPKM normalization on each bin. 'Same strand' and 'opposite strand' refer to DNA:RNA hybrid and ssDNA strands, respectively.

In order to compare the ssDRIP-seq peaks in terms of quality, we divided the peaks into ten subgroups considering their quality scores and tested the subgroups with the highest and the lowest quality scores in terms of damage distribution (Figure 4.19a, b). In addition, we included 10 kb-sized random regions from the genome to compare the damage abundance with the actual ssDRIP-seq peaks. Abundances of both damage types were higher on both DNA:RNA hybrid and ssDNA strand of 'top10' ssDRIP-seq peaks than the DNA:RNA hybrid and ssDNA strand of 'bottom10' ssDRIP-seq peaks, respectively (Figure 4.19a, b). On the other hand, the damage distribution on both subgroups showed the same pattern, being higher on DNA:RNA hybrid and lower on ssDNA around the peak centers than their flanking sites. In addition, the damage abundance was steady throughout the 10 kb length of the random regions and it was lower than the same abundance on both strands of the actual ssDRIP-seq peaks.

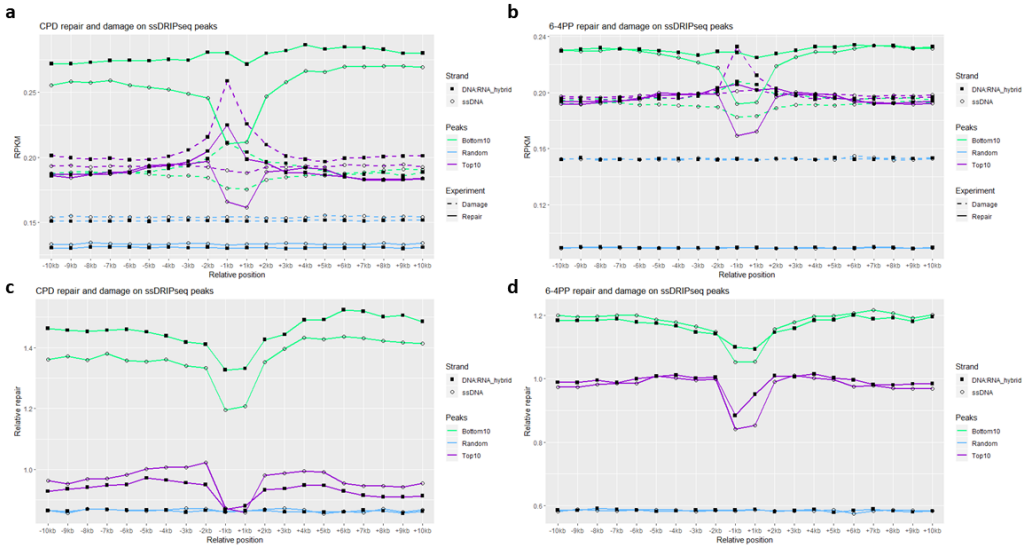

Figure 4.19 Damage-seq and XR-seq read distributions and relative repair rates on ssDRIP-seq R-loops in HeLa cells. Damage-seq (0-hour time-point) and XR-seq (12-minute time-point) in HeLa cells were used to obtain read distributions and relative repair rates on ssDRIP-seq R-loops centers and -/+ 10 kb flanking sites. (a) Damage and repair profiles of CPD lesions. 20 kb regions were divided into 20 bins and Damage-seq and XR-seq read counts were subjected to RPKM normalization on each bin. 'Same strand' and 'opposite strand' refer to DNA:RNA hybrid and ssDNA strands, respectively. Peaks with top and bottom 10% quality scores and random regions from the genome were included in the analysis and plotted separately. (b) Damage and repair profiles of (6-4)PP lesions. Damage and repair abundances were calculated as in (a). (c) CPD relative repair profiles on ssDRIP-seq R-loops centers and -/+ 10 kb flanking sites and random regions. Relative repair rates were calculated by normalizing the XR-seq RPKM with the Damage-seq RPKM in each window. (d) (6-4)PP relative repair profiles on ssDRIP-seq R-loops centers and -/+ 10 kb flanking sites and random regions. The relative repair rates were calculated as in (c).

The second data we processed was the qDRIP-seq data obtained from HeLa cells (Crossley et al., 2020). We aimed to check the CDP and (6-4)PP distribution on qDRIP-seq R-loop centers and 10 kb flanking regions upstream and downstream of the centers and compare the damage distribution profiles between the R-loop peaks from different methods. We also created simulated Damage-seq data in order to assess and eliminate the dependency of damage formation tendency to sequence content. The CPD and (6-4)PP levels on qDRIP-seq R-loops were similar throughout the 20 kb-sized regions with both damage types (Figure 4.20, left panels). At the R-loop centers, the abundance of both damage types peaked on DNA:RNA hybrid strand while damage levels were similar on ssDNA between centers and flanking sites. Simulated Damage-seq abundances were at comparable levels with the real

Damage-seq read abundances while being slightly higher on R-loop centers than the flanking regions (Figure 4.20, left panels). On the flanking regions, CPD distribution showed a strand difference being higher on the flanking sites of ssDNA (Figure 4.20, lower left panel). This difference was also observed with the simulated data (Figure 4.20, lower right panel), indicating that these regions might have a tendency to form CPD lesions due to their sequence contents.

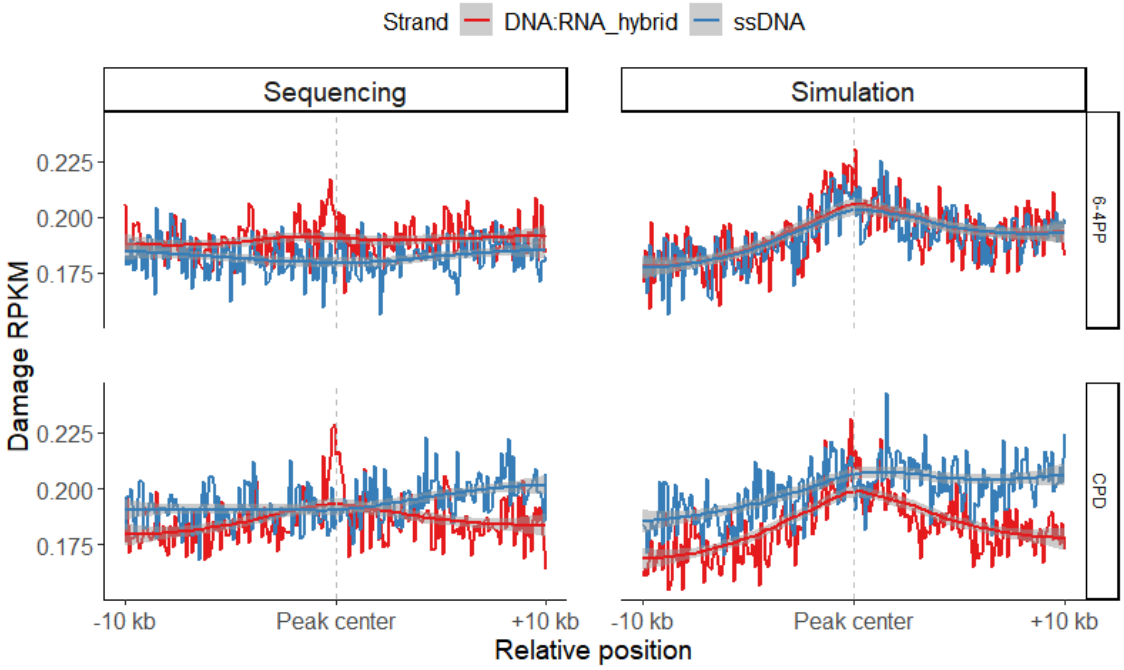

Figure 4.20 Damage-seq and simulated Damage-seq read distributions on qDRIP-seq R-loops in HeLa cells. Damage-seq (0-hour time-point) and simulated Damage-seq data of HeLa cells were used to obtain read distributions on qDRIP-seq R-loops centers and -/+ 10 kb flanking sites. 20 kb regions were divided into 20 bins and Damage-seq read counts were subjected to RPKM normalization on each bin. Damage-seq and simulated Damage-seq read abundances for CPD and (6-4)PP damage types were plotted separately.

RR-ChIP-seq R-loops showed a completely different damage accumulation pattern than qDRIP-seq peaks (Figure 4.20, 4.21). Damage-seq and simulated Damage-seq data of both damage types were lower on R-loop centers than on flanking sites on both DNA:RNA hybrid and ssDNA strands (Figure 4.21, left panels) while it was higher on DNA:RNA hybrid strand of qDRIP-seq peaks (Figure 4.20, left panels). An even more significant drop in damage abundance on RR-ChIP-seq R-loop centers was observed with the simulated Damage-seq data (Figure 4.21, right panels). On the other hand, the strand difference in damage abundance on the flanking sites was consistent with the strand difference on qDRIP-seq peaks (Figure 4.21, 4.20, lower panels). Interestingly, damage distributions of neither of the two data were

consistent with the distribution profile on ssDRIP-seq R-loops (Figure 4.19).

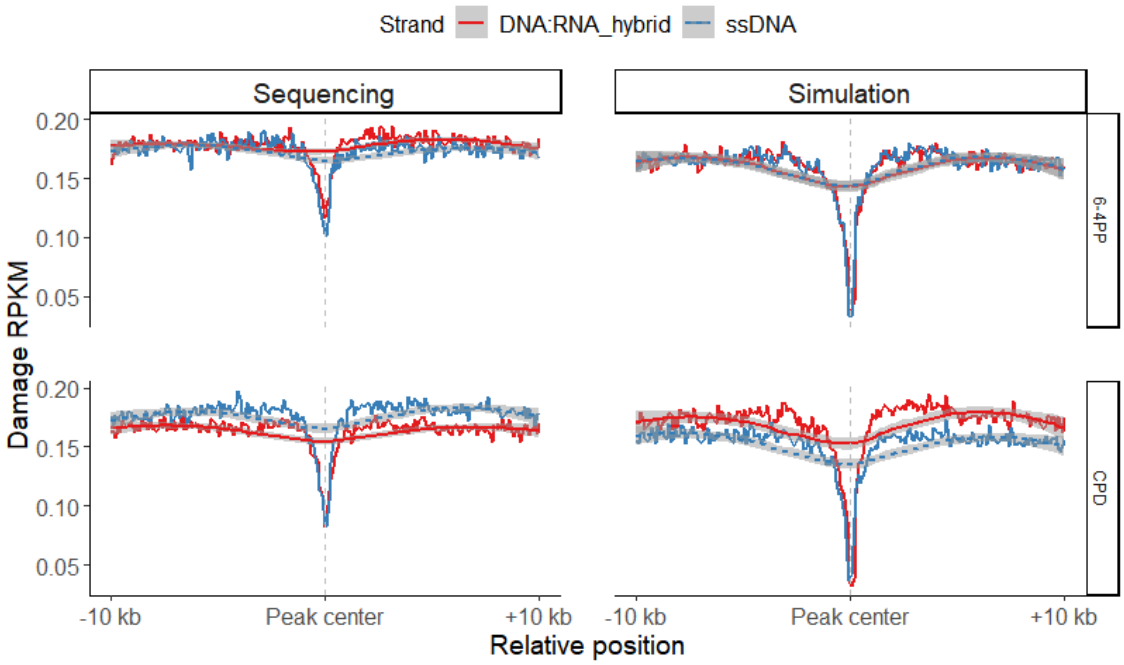

Figure 4.21 Damage-seq and simulated Damage-seq read distributions on RR-ChIP-seq R-loops in HeLa cells. Damage-seq (0-hour time-point) and simulated Damage-seq data of HeLa cells were used to obtain read distributions on RR-ChIP-seq R-loops centers and -/+ 10 kb flanking sites. 20 kb regions were divided into 20 bins and Damage-seq read counts were subjected to RPKM normalization on each bin. Damage-seq and simulated Damage-seq read abundances for CPD and (6-4)PP damage types were plotted separately.

In order to eliminate the dependency on sequence content, we have normalized the Damage-seq read abundances with the simulated Damage-seq read abundances and obtained the 'normalized damage' rates. Normalized damage profiles on qDRIP-seq and RR-ChIP-seq R-loops showed a completely opposite pattern for both CPD and (6-4)PP damage types (Figure 4.22, 4.23). On qDRIP-seq R-loops, damage on R-loop centers were lower than the flanking regions (Figure 4.22). Oppositely, damage on RR-ChIP-seq R-loop centers were much higher than their flanking regions (Figure 4.23).

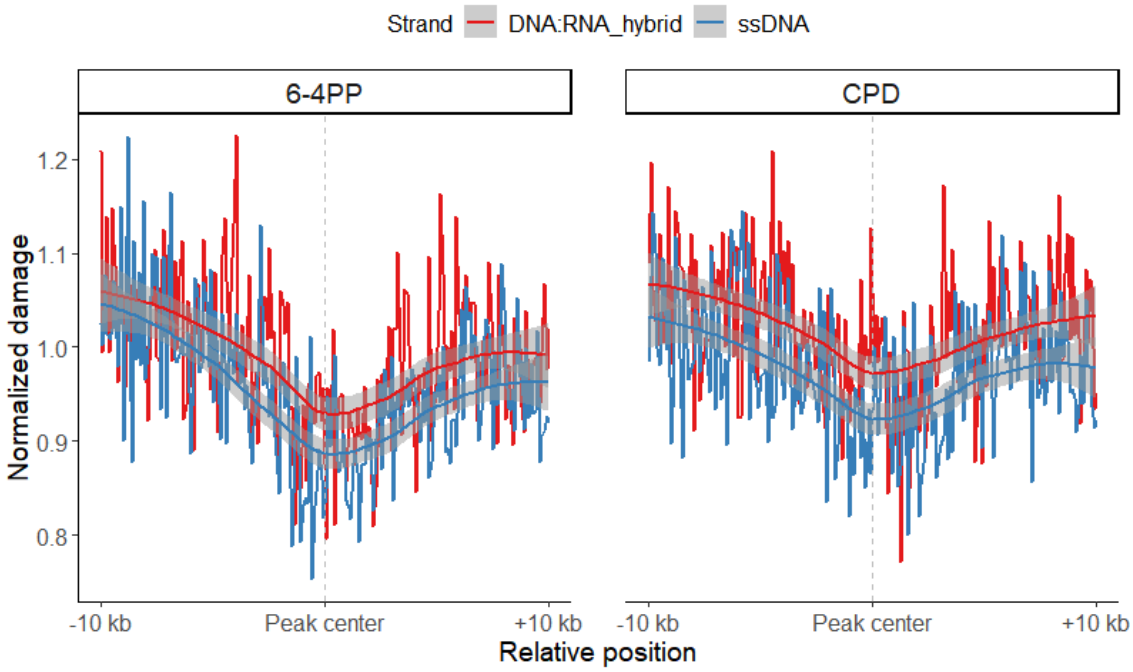

Figure 4.22 Normalized damage rates on qDRIP-seq R-loops in HeLa cells. Damage-seq (0-hour time-point) and simulated Damage-seq read abundances on qDRIP-seq R-loops centers and -/+ 10 kb flanking sites were counted and subjected to RPKM normalization on each bin. Damage-seq RPKMs were normalized with simulated Damage-seq RPKMs to obtain normalized damage rates. Normalized damage rates for CPD and (6-4)PP damage types were plotted separately.

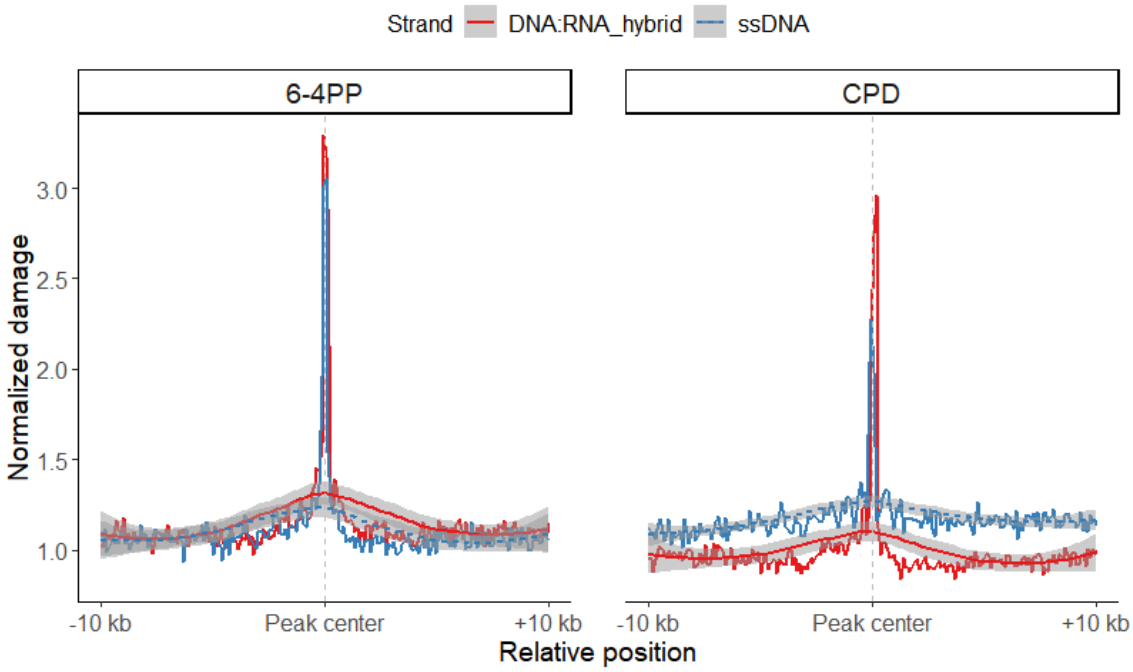

Figure 4.23 Normalized damage rates on RR-ChIP-seq R-loops in HeLa cells. Damage-seq (0-hour time-point) and simulated Damage-seq read abundances on RR-ChIP-seq R-loops centers and -/+ 10 kb flanking sites were counted and subjected to RPKM normalization on each bin. Damage-seq RPKMs were normalized with simulated Damage-seq RPKMs to obtain normalized damage rates. Normalized damage rates for CPD and (6-4)PP damage types were plotted separately.

The contrasting damage profiles as well as the differences in histone marker distributions (Figure 4.9, 4.10), chromatin accessibility levels (Figure 4.8) and genome coverage (Figure 4.3) between ssDRIP-seq, qDRIP-seq and RR-ChIP-seq R-loops led us to question the quality of these datasets and the accuracy of the three methods. In addition, due to the dynamic natures of the R-loops and the potential differences between the antibody-based and RNase H-based methods, these three data might be representing different sets of R-loops that are present on the genome at different times and on various regions.

While different methods sequenced inconsistent R-loop sets, R-loopBase database, published in 2022, curated all human R-loops sequenced by various methods and provided the common ones in nine levels considering the number of data that each R-loop position was found in (Lin et al., 2022). Of these levels, level 1 included the R-loops that were found in at least one of the data while level 9 R-loops were found in at least nine datasets. As being found in more datasets could correlate with the accuracy of that R-loop position, we choose levels 7, 8 and 9 for the further analyzes. However, the number of level 9 R-loops were very low; therefore, we continued with level 7 and 8 R-loops.

The normalized damage profiles on level 7 (Figure 4.24a) and level 8 (Figure 4.24b) R-loops were similar to the profiles on RR-ChIP-seq R-loops (Figure 4.23). The abundance of both CPD and (6-4)PP damages were higher on R-loops centers than on flanking sites. No strand difference was observed on the flanking regions while on the centers, ssDNA accumulated higher damage than DNA:RNA hybrid, which was contrasting with the RR-ChIP-seq R-loops. Despite R-loopBase database provided R-loops shared by several data from different methods, the results were different than the ones we obtained in the previous analyzes. Therefore, to be more confident, we kept searching for other methods and datasets to assess R-loops.

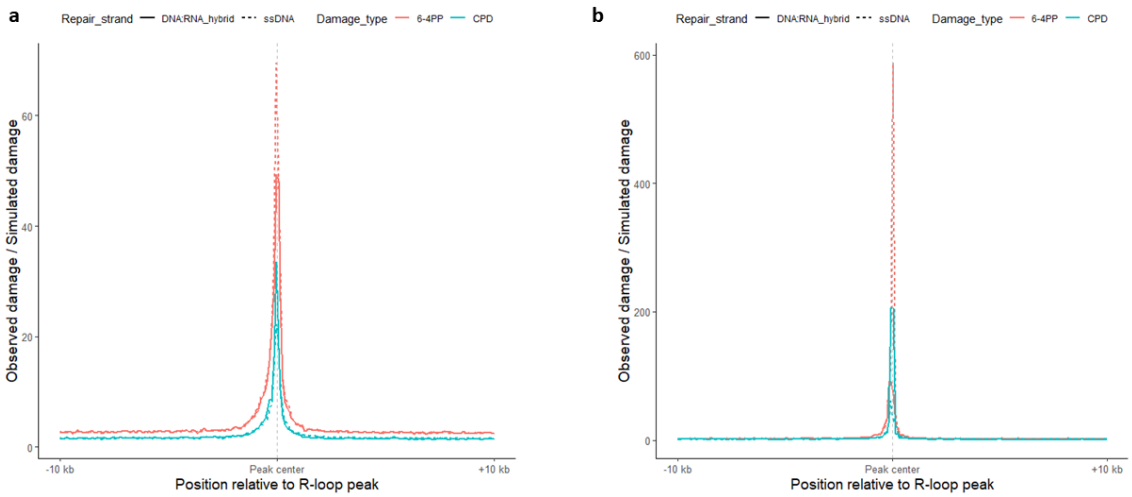

Figure 4.24 Normalized damage rates on R-loopBase R-loops in HeLa cells. Damage-seq (0-hour time-point) and simulated Damage-seq read abundances on R-loopBase R-loops centers and -/+ 10 kb flanking sites were counted and subjected to RPKM normalization on each bin. Damage-seq RPKMs were normalized with simulated Damage-seq RPKMs to obtain normalized damage rates. Normalized damage rates for CPD and (6-4)PP damage types were plotted separately.

When the RLBase database was published which performed a quality-control assessment on the human R-loops in the literature and provided the consensus R-loop sites (Miller et al., 2022), we have examined the damage distribution on RLBase R-loops. Similar to the analysis with the previously published R-loop sites, we have calculated normalized damage on each region by dividing the RPKM-normalized Damage-seq counts with the RPKM-normalized simulated Damage-seq counts. We have performed these analyzes using Damage-seq data from NHF1 (Hu et al., 2017) and HeLa (Huang et al., 2022) cells. Using the HeLa Damage-seq data obtained right after the UV exposure (0-hour time-point), we observed that the damage accumulation pattern around R-loop centers were contrasting between the two strands (Figure 4.25). On DNA:RNA hybrid strand, a higher damage accumulation was observed than on ssDNA and dsDNA at the flanking sites while the damage accumulation

on ssDNA was the lowest among the three structures. Even though the damage on DNA:RNA hybrid was high on a 2 kb region spanning the R-loop centers, a decrease right at the R-loop centers was observed. This damage peak on DNA:RNA hybrid and drop on ssDNA was consistent with the length distribution of R-loops as they mostly covered -/+ 2 kb regions around R-loop centers (Figure 4.17).

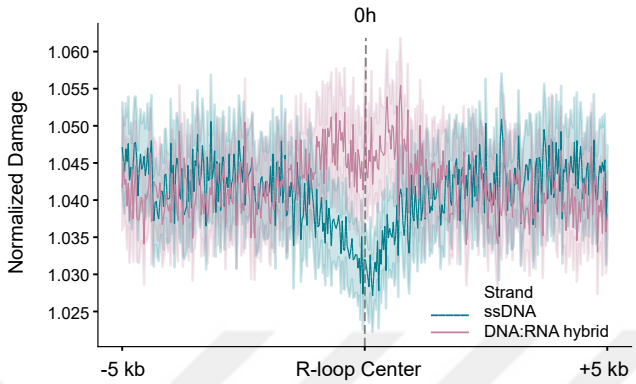

Figure 4.25 CPD damage distribution on RLBase R-loops in HeLa cells. CPD Damage-seq (0-hour time-point) and simulated Damage-seq data of HeLa cells were used to obtain damage distributions on RLBase R-loop centers and -/+ 5 kb flanking sites. 10 kb regions were divided into 400 bins and Damage-seq and simulated Damage-seq read counts were subjected to RPKM normalization. Damage-seq RPKM was normalized by simulated Damage-seq RPKM to obtain normalized damage on each bin.

We observed the same CPD damage distribution when we used the Damage-seq data obtained from NHF1 cells at the initial time-point after UV exposure (0h) (Figure 4.26). At the 1-hour time point, the profile resembled the 0-hour damage profile while the levels of damage slightly dropped as the repair mechanism eliminated the damaged sites. At the 8-hour time point, the damage level drop on DNA:RNA hybrid and its flanking sites was more explicit than the drop on ssDNA. This could be due to the preferential activity of TC-NER on TS of genes which becomes fully functional later than the GG-NER mechanism (Hu et al., 2017). Since R-loops are generally produced during transcription, their DNA:RNA hybrid strands form on the TS which is more efficiently repaired by TC-NER after the first hour of UV exposure (Castillo-Guzman & Chédin, 2021; Hu et al., 2017). Despite this significant drop on DNA:RNA hybrid, ssDNA damage levels and profile remained similar. A higher drop on ssDNA damage levels was observed at the 24-hour time-point, indicating that the repair on ssDNA and therefore NTS was subjected to a delay.

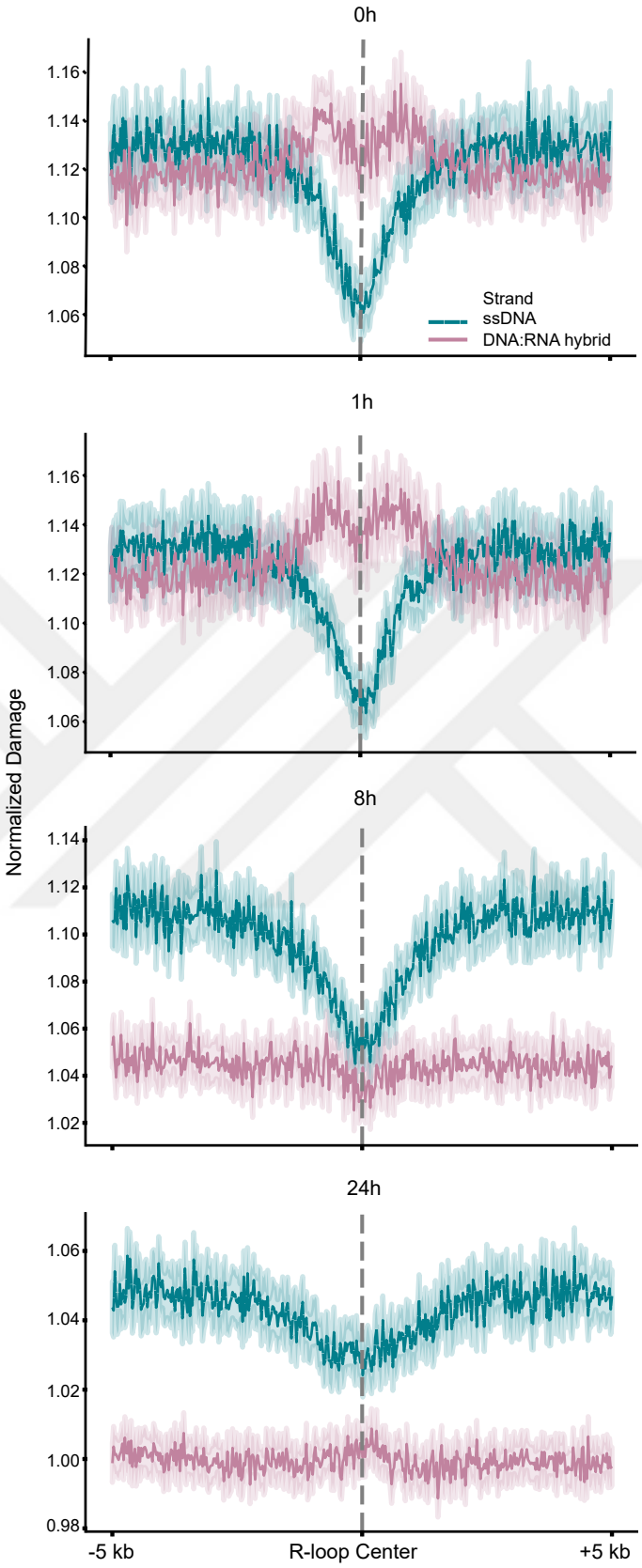

Figure 4.26 CPD damage distribution on RLBase R-loops in NHF1 cells. CPD Damage-seq (0, 1, 8, 24-hour time-points) and simulated Damage-seq data of NHF1 cells were used to obtain damage distributions on RLBase R-loop centers and -/+ 5 kb flanking sites. 10 kb regions were divided into 400 bins and Damage-seq and simulated Damage-seq read counts were subjected to RPKM normalization. Damage-seq RPKM was normalized by simulated Damage-seq RPKM to obtain normalized damage on each bin.

Comparing gene segments that contained R-loop DNA:RNA hybrids on their TS, which is the common orientation, and gene segments that did not contain any R-loops in terms of CPD accumulation revealed significant differences between TS and NTS of genes with and without R-loops (Figure 4.27). On TS, genes with R-loops accumulated significantly higher damage whereas on NTS, this pattern was the opposite. There results were consistent with the general CPD profiles on R-loops (Figure 4.25, 4.26).

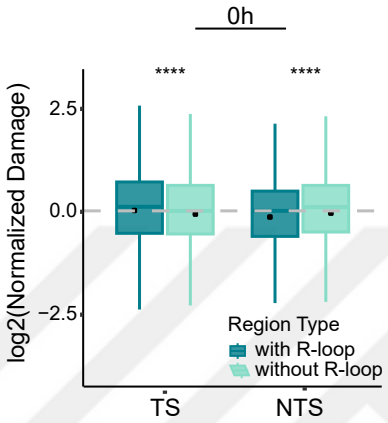

Figure 4.27 CPD damage distribution differences between gene segments overlapping with R-loops and other gene segments. Genes were overlapped with R-loop centers and spanning 2 kb regions and gene segments with overlaps were defines as 'with R-loops'. 'without R-loops' class of gene segments were randomly selected from genes with no overlap with any whole R-loop regions. Normalized damage was calculated as explained in Figure 4.25 using Damage-seq data (0-hour) obtained from NHF1 cells. t-test was used to compute P-values. (ns: $p > 0.05$; *: $p <= 0.05$; **: $p <= 0.01$; ***: $p <= 0.001$; ****: $p <= 0.0001$)

To understand more about the reasons behind the differential damage levels on DNA:RNA hybrid and ssDNA structures, we have checked if the damage accumulation differed when there was G-quadruplex (G4) formation on R-loop ssDNA strand. G4s are another type of secondary DNA structures that are formed when one of the DNA strands folds and becomes four-stranded and stabilizes itself with hydrogen bonds (Monsen, Trent & Chaires, 2022). Comparison of R-loops that contained a G4 on their ssDNA and R-loops that did not intersect with G4s at neither strand revealed a significantly lower damage formation on both strands of R-loops with a G4 on ssDNA (Figure 4.28), indicating that G4 presence may affect damage formation even though it did not explain the difference between the two strands of R-loops.

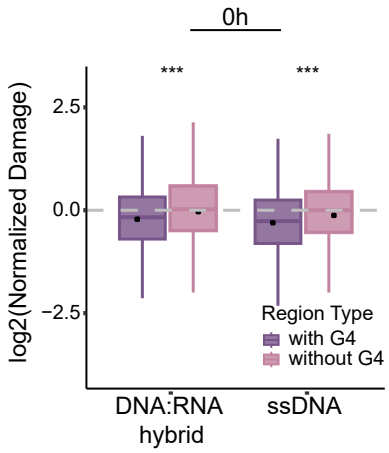

Figure 4.28 CPD damage distribution differences between R-loops overlapping with G4 structures and other R-loops. R-loop ssDNA strands were overlapped with G4 sites and defines as 'with G4s' if overlap was detected. 'without G4s' class of R-loops were randomly selected from R-loops with no overlap with any G4s. Normalized damage was calculated as explained in Figure 4.25. t-test was used to compute P-values. (ns: p > 0.05; *: p <= 0.05; **: p <= 0.01; ***: p <= 0.001; ****: p <= 0.0001)

The damage distribution results on RLBase R-loops suggested that the R-loop regions alter the CPD formation potentials of the DNA regions. Importantly, this potential was different of ssDNA and DNA:RNA hybrid strands. The results indicated that ssDNA lowers the tendency of CPD formation as opposed to the DNA:RNA hybrid with higher damage accumulation than the duplex DNA at the flanking sites of R-loops.

In conclusion, we have assessed the damage formation patterns on R-loops from different methods and databases. After careful comparison, we realized the inconsistencies between the methods and decided that using R-loops from a database which assessed R-loop sites with a set of quality criteria would be more reliable. Therefore, we used RLBase R-loops for the further analyzes.

4.2.1.1 Molecular dynamics simulations of ssDNA, DNA:RNA hybrid and dsDNA structures

The formation of CPD is generally influenced by the spatial separation and rotational orientation of the C5-C6 double bonds in the aromatic rings of neighboring pyrimidines. The production of CPD is increasingly favorable when the distance

and angle between these bonds decrease (Law et al., 2008; Nayis et al., 2023; Stark et al., 2022). Due to our previous observations that DNA:RNA hybrid accumulated higher CPD damage than the double-stranded DNA (dsDNA) and the ssDNA, and single-stranded DNA (ssDNA) had the lowest level of CPD damage among the three structures, we conducted molecular dynamics (MD) simulations to investigate the variations in dipyrimidine distances (Figure 4.29) and angles (Figure 4.30) among three configurations: ssDNA, DNA:RNA hybrid, and dsDNA. The molecular dynamics (MD) simulations were conducted for each of the three molecules utilizing the crystal structure of a DNA:RNA hybrid molecule (PDB ID: 1G4Q). This particular hybrid molecule contained consecutive thymines (Ts) inside its DNA sequence. In order to facilitate a comparative analysis of the behaviors shown by adjacent Ts inside three distinct structures, we initiated simulations of both ssDNA and dsDNA using the DNA structure in the crystal structure of the DNA:RNA hybrid. Initially, we observed a hairpin formation at the 3' end of the ssDNA structure and therefore, we restricted the movement of the nucleotides at the 5' and 3' ends. We used Ts at the 2nd and 3rd, and 7th and 8th positions in the structures to measure the distances and angles.

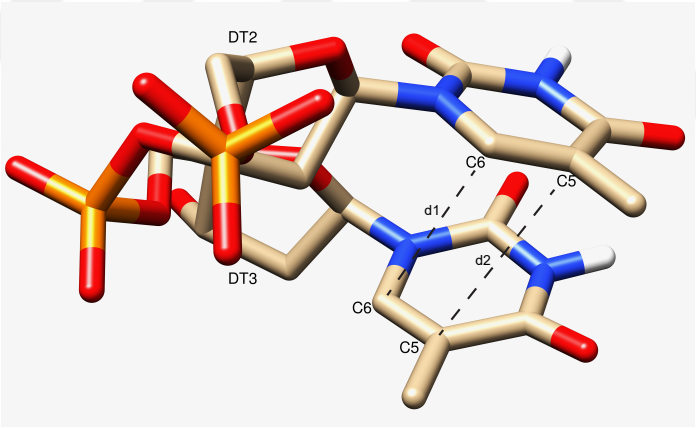

Figure 4.29 Distance measurements between adjacent thymines. The distances between the C5s and C6s of adjacent thymines were measured.

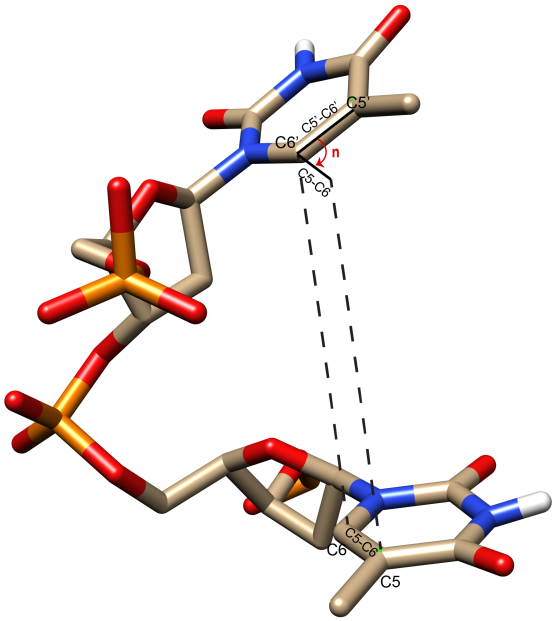

Figure 4.30 Angle measurements between C5-C6 bonds adjacent thymines. The angles between the C5-C6 bonds in the aromatic rings of adjacent thymines were measured.

The distance measurements between the C5s and C6s of the adjacent Ts revealed that the mean distances calculated from the measurements taken at every 20 picoseconds (ps) during a 400-nanosecond (ns) trajectory were lower than 5 Å between both 2nd and 3rd, and 7th and 8th TT pairs in all three molecules, which indicated that in terms of distance, these TT pairs in the three molecules hold the potential the form CPDs upon UV exposure (Figure 4.31). On the other hand, remarkable distance fluctuations were observed between the TT pairs of ssDNA structure whereas the measured distances in dsDNA and DNA:RNA hybrid were more stable (Figure 4.32). These fluctuations were even more notable between the TT pair in 7th-8th positions which could be due to this pair being a little further from the molecule ends where there are movement restrictions in the MD simulations, when compared to the other TT pair. Being closer to the molecule ends might have created an obstacle for the free movement at the 2nd and 3rd positions whereas the nucleotides could be more motile at the 7th and 8th positions. In summary, not the mean distances but the more motile behaviors of the nucleotides in ssDNA due to lacking an opposite strand which could hold the nucleotides and restrict their movements, could be one of the reasons why we observed lower normalized CPD damage on the ssDNA. More specifically, the results suggested that the pyrimidines in the ssDNA were so motile that they rarely come to closer distances with each other to make CPD formation possible. In addition, the distance measurements being more stable in DNA:RNA hybrid than in dsDNA could explain the normalized CPD damage being higher on

DNA:RNA hybrid with the increment in the CPD formation possibility as the C5-C6 bonds in the adjacent pyrimidines are mostly closer to each other.

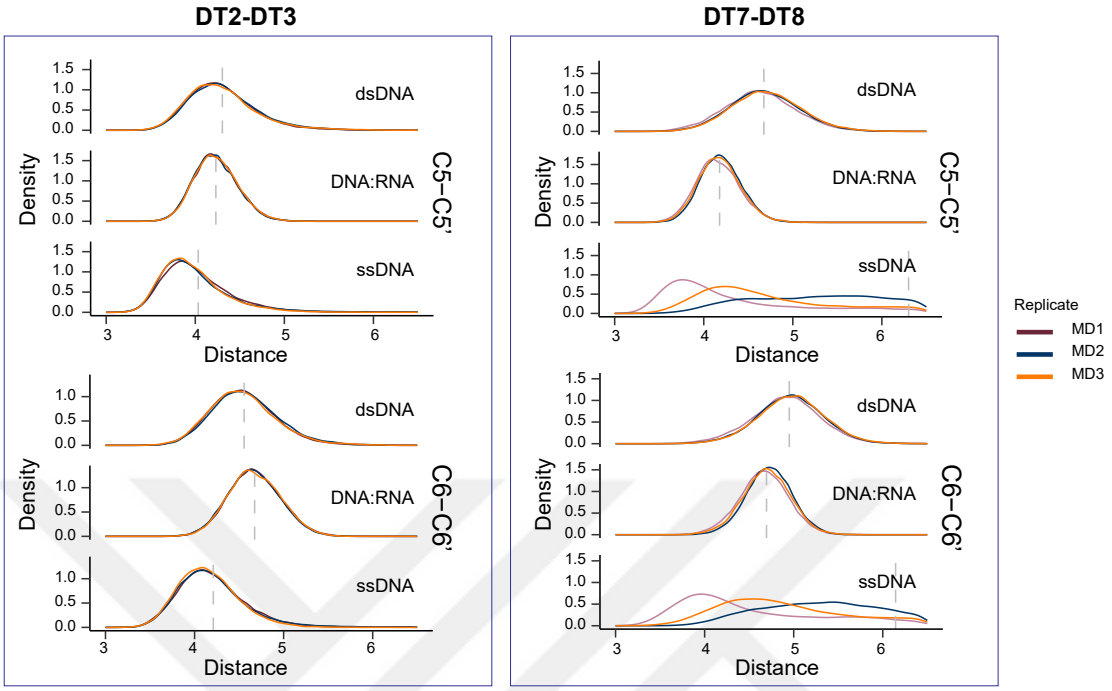

Figure 4.31 Density distributions of the distances between C5s and C6s of the adjacent thymines in dsDNA, DNA:RNA hybrid and ssDNA structures. The MD simulations for the three structures were run as three replicates through a 500-nanosecond (ns) interval. The first 100 ns was not included in the measurements. The distances between the C5s and C6s of adjacent thymines were measured at every 20 picoseconds (ps) in all three structures. Two TT pairs at 2nd and 3rd (DT2-DT3), and 7th and 8th (DT7-DT8) positions were measured throughout the trajectory. Density plots were generated for each molecule and replicate separately using ggplot (Wickham, 2011). The mean distances including the measurements from the three replicates were shown with the dashed lines.

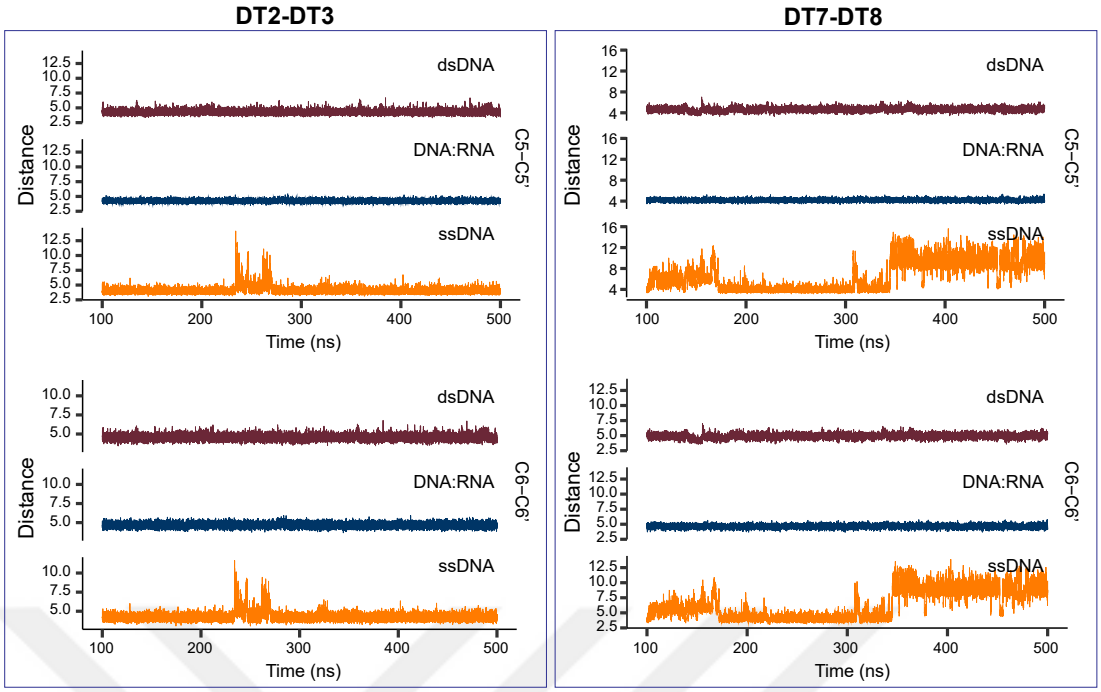

Figure 4.32 Distances between C5s and C6s of the adjacent thymines in dsDNA, DNA:RNA hybrid and ssDNA structures. The MD simulations for the three structures were run through a 500-nanosecond (ns) interval. The first 100 ns was not included in the measurements. The distances between the C5s and C6s of adjacent thymines were measured at every 20 picoseconds (ps) from the first replicates of all three structures. Two TT pairs at 2nd and 3rd (DT2- DT3), and 7th and 8th (DT7-DT8) positions were measured throughout the trajectory.

The mean angles between the C5-C6 double bonds of the adjacent Ts differentiated from each other more than the distances in the three structures. The lowest means angles were measured in the DNA:RNA hybrid and the highest mean angles were measured in ssDNA (Figure 4.33a). The angle differences between the three molecules were consistent with the damage levels on them considering the phenomena that lower angles favor the CPD formation (Mao et al., 2018; Nayis et al., 2023). The lowest normalized CPD levels were detected on the ssDNA which had the highest mean angles between both TT pairs, while the highest normalized CPD levels were on the DNA:RNA hybrid where the lowest mean angles were measured (Figure 4.25, 4.26). Moreover, similar to the distance fluctuations, the angle fluctuations were also observed in ssDNA structure which could be an additional factor restricting the CPD formation (Figure 4.33b). Altogether, we speculated that the reason for the reduced CPD damage on the ssDNA could be the high motility of the nucleotides which rarely align in the proper position with respect to each other in order for the CPD formation reaction can occur successfully. In addition, the strong interactions between the DNA and RNA strands might be keeping the nucleotides in a less motile form which may lead to increase in CPD formation potential as the

adjacent pyrimidines are mostly in the close distance and aligned laterally.

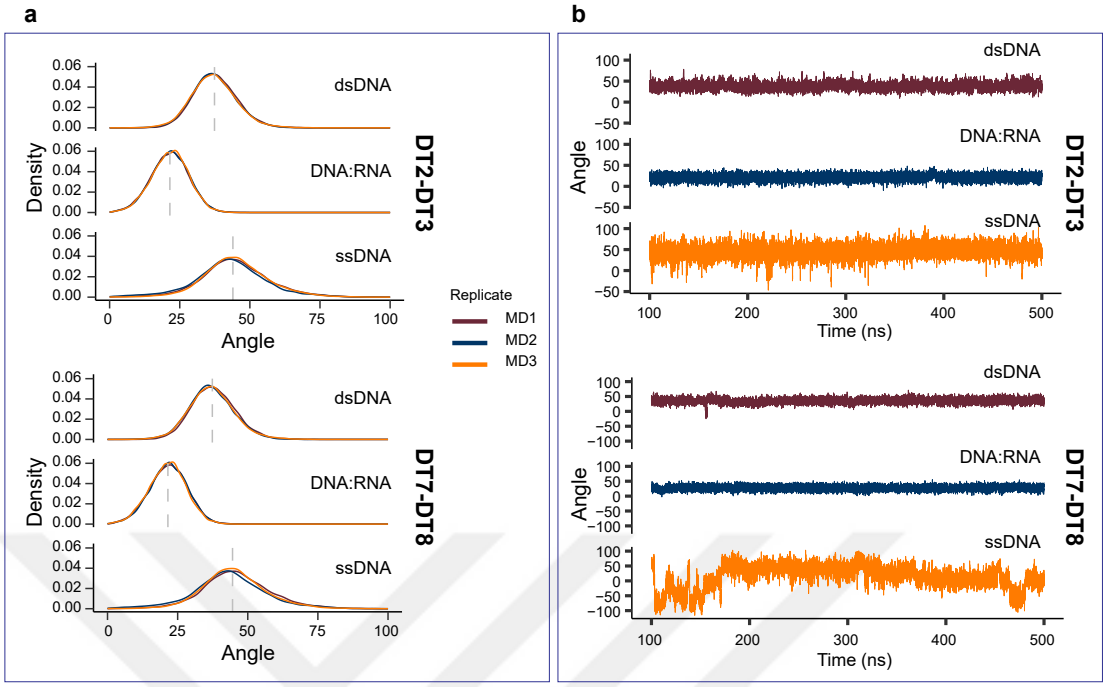

Figure 4.33 Angles between the C5-C6 double bonds of adjacent thymines in ds-DNA, DNA:RNA hybrid and ssDNA structures. The angles between the C5-C6 double bonds of adjacent thymines were measured in the three structures. The MD simulations for the three structures were run through a 500-nanosecond (ns) interval. The first 100 ns was not included in the measurements. The distances between the C5s and C6s of adjacent thymines were measured at every 20 picoseconds (ps) in all three structures. Two TT pairs at 2nd and 3rd (DT2- DT3), and 7th and 8th (DT7-DT8) positions were measured throughout the trajectory. (a) Density distributions of the measured angles. The mean distances including the measurements from the three replicates were shown with the dashed lines. (b) Angle measurements from the replicate 1 throughout the 400-ns trajectory. Plots were generated using ggplot (Wickham, 2011).

4.2.2 Repair profiles on R-loops

Similar to damage formation tendency, nucleotide excision repair (NER) efficiency on R-loops has never been established clearly. R-loops are prevalent structures on our genomes and therefore, it is very important to understand how capable NER mechanism is in repairing the UV-induced lesions on ssDNA and DNA:RNA hybrid structures of R-loops. To analyze this, we used XR-seq datasets which provided regions on the genome where NER mechanism was active by capturing the excision products released from the damaged sites when the endonucleases in NER pathway

cut the damaged DNA strand from the two sides of the damage (Hu et al., 2015; Hu, Li, Adebali, Yang, Oztas, Selby & Sancar, 2019). As we did for the damage profiles, we analyzed and compared repair on R-loops coming from three methods, ssRIP-seq, qDRIP-seq and RR-ChIP-seq, and two databases, R-loopBase and RLBase.

Firstly, we checked ssDRIP-seq R-loops for repair using the XR-seq and Damage-seq data obtained from HeLa cells after UV exposure (Huang et al., 2022). We observed a higher CPD repair on DNA:RNA hybrid when compared to the ssDNA and the flanking sites while the repair on ssDNA was much lower than DNA:RNA hybrid and the flanking regions (Figure 4.18a). Even though the repair on ssDNA was similar between CPD and (6-4)PP, unlike CPD repair, the (6-4)PP repair on DNA:RNA hybrid was not so different from its flanking sites (Figure 4.18b). Repair of both types of damages were higher on the ssDRIP-seq R-loops than the random regions (Figure 4.19a, b).

The intensity of the repair activity on a particular region on the genome depends on the number of the damaged sites on that region. Thus, when we detect a region where the repair activity is low, it is hard to understand if the reason for low repair is the low damage abundance or other factors that restrict the repair activity. To overcome this, we have normalized repair by the damage on the same region to eliminate the effect of damage abundance on the repair activity. After this normalization, we observed lower repair on both strands of R-loops when compared to their surrounding regions (Figure 4.19c, d).

To eliminate another dependency, we have produced simulated XR-seq and Damage-seq data sets using Boquila tool (Akkose & Adebali, 2023). We have obtained the 'normalized repair' rates by normalizing the read abundances of XR-seq data with the read abundances in simulated XR-seq data on each genomic region in the analysis. By repeating the same normalization with the Damage-seq and simulated Damage-seq data, we have obtained the 'normalized damage' ratios. Finally, by dividing normalized repair by the normalized damage, we have calculated 'relative repair' rates which provided the repair efficiency that was independent of the damage abundance and the sequence content of the genomic region of interest.

After ssDRIP-seq R-loops, we assessed qDRIP-seq R-loops in terms of repair efficiency. We calculated the relative repair on qDRIP-seq R-loop centers and 10 kb upstream and downstream of the centers using the XR-seq and Damage-seq data obtained from HeLa cells after UV exposure (Huang et al., 2022). The relative repair profiles for (6-4)PP and CPD damages differed (Figure 4.34). (6-4)PP repair was more efficient on R-loop centers than the CPD repair while the repair levels were close to each other on the two strands on flanking sites (Figure 4.34, left panel). On

the other hand, CPD repair was more efficient on the DNA:RNA hybrid strand and its flanking sites than the ssDNA and its flanking sites (Figure 4.34, right panel). The repair peaked not at the R-loop peak centers, but at around 2 kb downstream of them. These profiles contradicted with the repair, normalized by the damage but not with the simulations, on ssDRIP-seq R-loops where the repair was lower on R-loop centers than their flanking regions (Figure 4.19c, d).

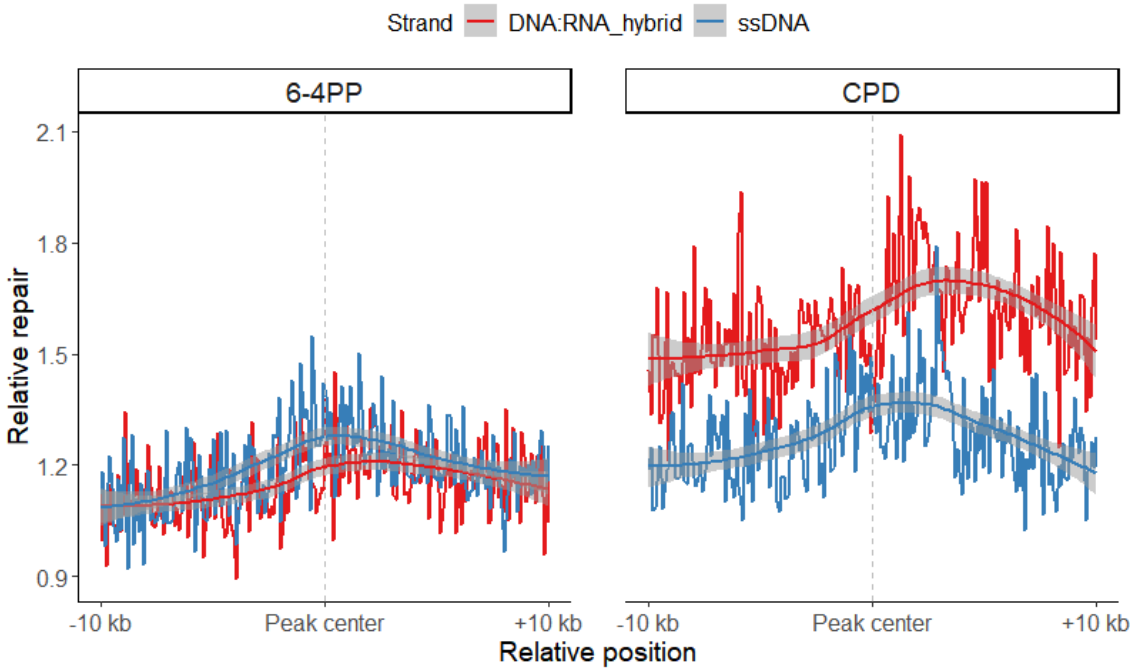

Figure 4.34 Relative repair profiles on qDRIP-seq R-loops in HeLa cells. Relative repair of CPD and (6-4)PP lesions on R-loop centers and -/+ 10 kb flanking regions. Regions were divided into 400 equal bins and relative repair rates were calculated by counting the Damage-seq (0-hour, HeLa), XR-seq (12-minute, HeLa) and simulated Damage-seq and XR-seq reads on each region. After RPKM normalizations, Damage-seq RPKMs were normalized by the simulated Damage-seq RPKMs to obtain normalized damage ratios for each region. The same thing was repeated with XR-seq and simulated XR-seq data to obtain normalized repair values. Finally, normalized repair on each region was divided by the normalized damage value to obtain relative repair rate.

The relative repair on RR-ChIP-seq R-loops also revealed contradicting results. The repair of (6-4)PP lesions was lower on R-loop centers than their flanking sites on both strands although it peaked on the adjacent genomic regions of the centers (Figure 4.35, left panel). CPD repair was higher on the regions adjacent to the R-loop centers while it dropped at the centers (Figure 4.35, right panel). In general, CPD repair on ssDNA was more efficient than on DNA:RNA hybrid strand.

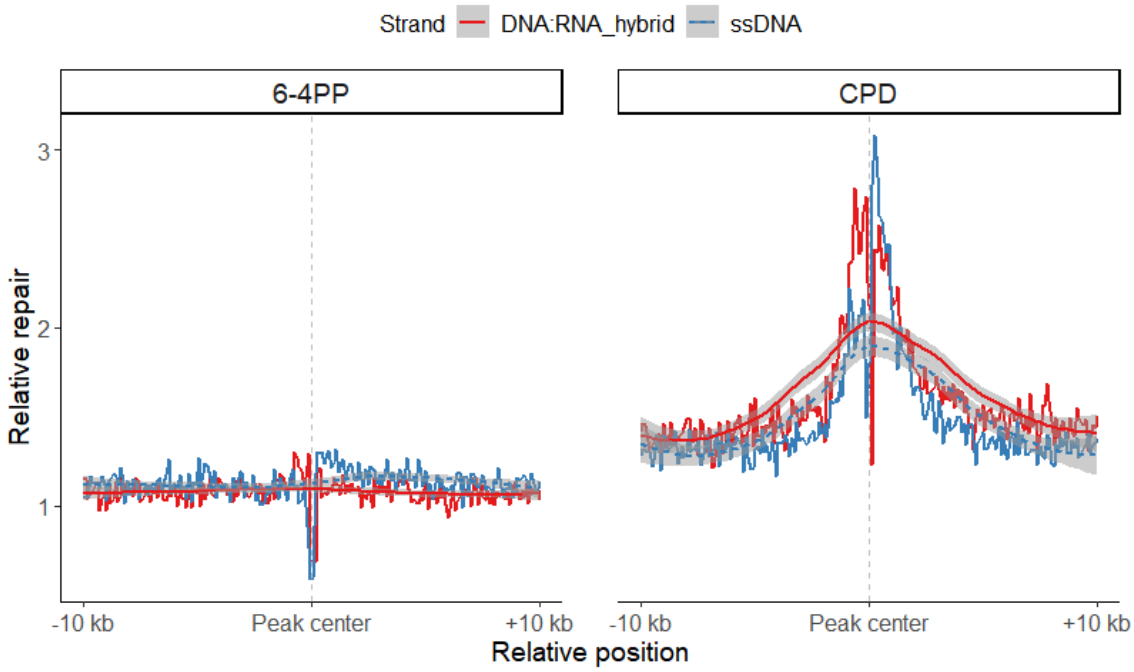

Figure 4.35 Relative repair profiles on RR-ChIP-seq R-loops in HeLa cells. Relative repair of CPD and (6-4)PP lesions on R-loop centers and -/+ 10 kb flanking regions. Regions were divided into 400 equal bins and relative repair rates were calculated by counting the Damage-seq (0-hour, HeLa), XR-seq (12-minute, HeLa) and simulated Damage-seq and XR-seq reads on each region. After RPKM normalizations, Damage-seq RPKMs were normalized by the simulated Damage-seq RPKMs to obtain normalized damage ratios for each region. The same thing was repeated with XR-seq and simulated XR-seq data to obtain normalized repair values. Finally, normalized repair on each region was divided by the normalized damage value to obtain relative repair rate.

The differences between the relative repair profiles of (6-4)PP and CPD damages were expected due to the structural differences between these two lesion types. Formation of (6-4)PP damage causes DNA to bend more than the formation of CPD and they are more efficiently repaired by NER mechanism (Hung, Sidorova, Nghiem & Kawasumi, 2020) which could explain the relative repair distinction between the two lesion types. However, the relative repair profile differences were not expected between the ssDRIP-seq, qDRIP-seq and RR-ChIP-seq R-loops since they represent the same cell line and ideally the same R-loop regions. Since we also observed inconsistencies between the genome coverages and damage profiles of the three R-loop data, we speculated that these datasets might have accuracy issues or they sequenced different sets of R-loops as R-loops tend to resolve and form dynamically at certain times depending on the transcription activities and the R-loop regulators (Castillo-Guzman & Chédin, 2021). Due to these reasons, we could not rely solely on any of these R-loop datasets and decided to continue with the consensus R-loop

regions from various human cell lines provided by the R-loopBase database (Lin et al., 2022).

The relative repair analysis of R-loopBase R-loops revealed a higher repair for CPD relative repair than (6-4)PP relative repair (Figure 4.36a). The (6-4)PP relative repair rates did not differ significantly between the R-loop centers and flanking regions while CPD relative repair was significantly higher on R-loop centers than on downstream flanking regions although the immediate upstream regions were also repaired efficiently (Figure 4.36b). When the two strands of R-loops were compared, ssDNA strand had a higher repair rate than the hybrid strand for CPD damage. In general, a peak for CPD repair at ssDNA was observed at the R-loop peak center. On the contrary, although CPD repair at DNA:RNA hybrid was higher at the R-loops than the flanking regions, there was a slight decline right at the R-loop peak center. Notably, this relative repair peak at ssDNA and decline at hybrid strand lied within zones where most of the R-loops covered (-/+ 300 bp from peak centers).

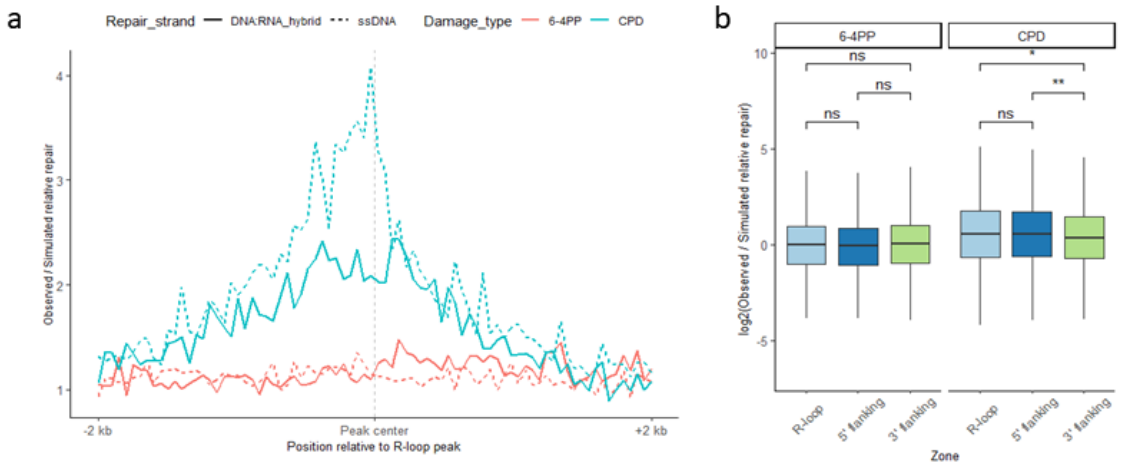

Figure 4.36 Relative repair profiles on R-loopBase R-loops in HeLa cells. (a) Relative repair of CPD and (6-4)PP lesions on R-loop centers and -/+ 2 kb flanking regions. Regions were divided into equal bins and relative repair rates were calculated by counting the Damage-seq (0-hour, HeLa), XR-seq (12-minute, HeLa) and simulated Damage-seq and XR-seq reads on each region. After RPKM normalizations, Damage-seq RPKMs were normalized by the simulated Damage-seq RPKMs to obtain normalized damage ratios for each region. The same thing was repeated with XR-seq and simulated XR-seq data to obtain normalized repair values. Finally, normalized repair on each region was divided by the normalized damage value to obtain relative repair rate. (b) Relative repair difference between R-loop centers and flanking regions. The flanking regions represent the zones adjacent to each R-loop with the same length as the R-loop. Relative repair was calculated on each region as explained in (a). The significance between relative repair rates was computed with t-test (ns: $p > 0.05$; *: $p <= 0.05$; **: $p <= 0.01$; ***: $p <= 0.001$; ****: $p <= 0.0001$)

To check the repair profiles on RLBase (Miller et al., 2022) R-loops, we have calculated the relative repair rates on RLBase R-loop centers and 10 kb regions spanning them as explained before using the Damage-seq and XR-seq data obtained from HeLa (12-minute time-point) (Huang et al., 2022) and NHF1 (1, 8, 24-hour time-points) cells (Hu et al., 2017), and the simulations. In HeLa cells, the results revealed that R-loop centers receive a more efficient repair then their flanking regions (Figure 4.37). The relative repair peak reached even higher on DNA:RNA hybrid while a slight drop was observed on both DNA:RNA hybrid and ssDNA peaks right at the R-loop centers. A strand difference was also observed at the flanking sites.

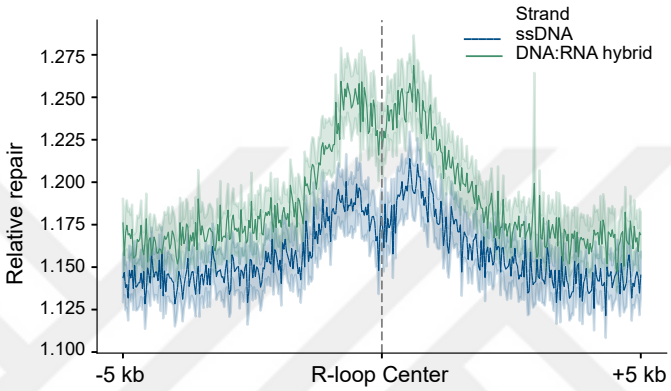

Figure 4.37 Repair profiles on RLBase R-loops in HeLa cells. R-loop centers and 5 kb upstream and downstream regions were taken into consideration and divided into 400 equal bins. Damage-seq (0-hour time-point) and XR-seq (12-minute time-point) data reads were intersected with the genomic bins and the overlaps were counted and subjected to RPKM normalization. The overlapped Damage-seq and XR-seq RPKMs were first normalized by simulated data RPKMs to obtain normalized damage and normalized repair, respectively. Then, normalized repair was divided by normalized damage on each bin to obtain relative repair.

Using the Damage-seq and XR-seq data obtained from NHF1 cells, we have observed similar relative repair profiles on RLBase R-loops. At 1-hour time-point, the relative repair on R-loop centers were repaired more efficiently than the flanking regions while there was a drop right at the peak centers (Figure 4.38) as observed in HeLa cells (Figure 4.37). In NHF1 cells (1-hour time-point), the level of relative repair was higher on DNA:RNA hybrid and its flanking sites than that of in HeLa cells (12-minute time-point), due to the further initiation of TC-NER at the end of the first hour. Due to the same reason, a more pronounced strand difference was observed in NHF1 cells. At the 8-hour time-point, the strand difference did not change significantly and the repair levels were similar to the ones at the first hour. At the final hour, relative repair on ssDNA was higher than on DNA:RNA hybrid and the relative repair on R-loops were lower than their surrounding sites, probably because the damaged sites were repaired on those regions before this time-point. In general,

this time-point data suggested that the strand difference might be due to the TC-NER activity on DNA:RNA hybrids which are mostly located on the TS of genes where TC-NER preferentially repairs (Hu et al., 2017). On the other hand, the higher relative repair on R-loop centers on both strands cannot be fully explained by the TC-NER activity since no TC-NER is expected on ssDNAs which are mostly located at the NTS of genes (de Laat, Jaspers & Hoeijmakers, 1999b; Reardon & Sancar, 2005).

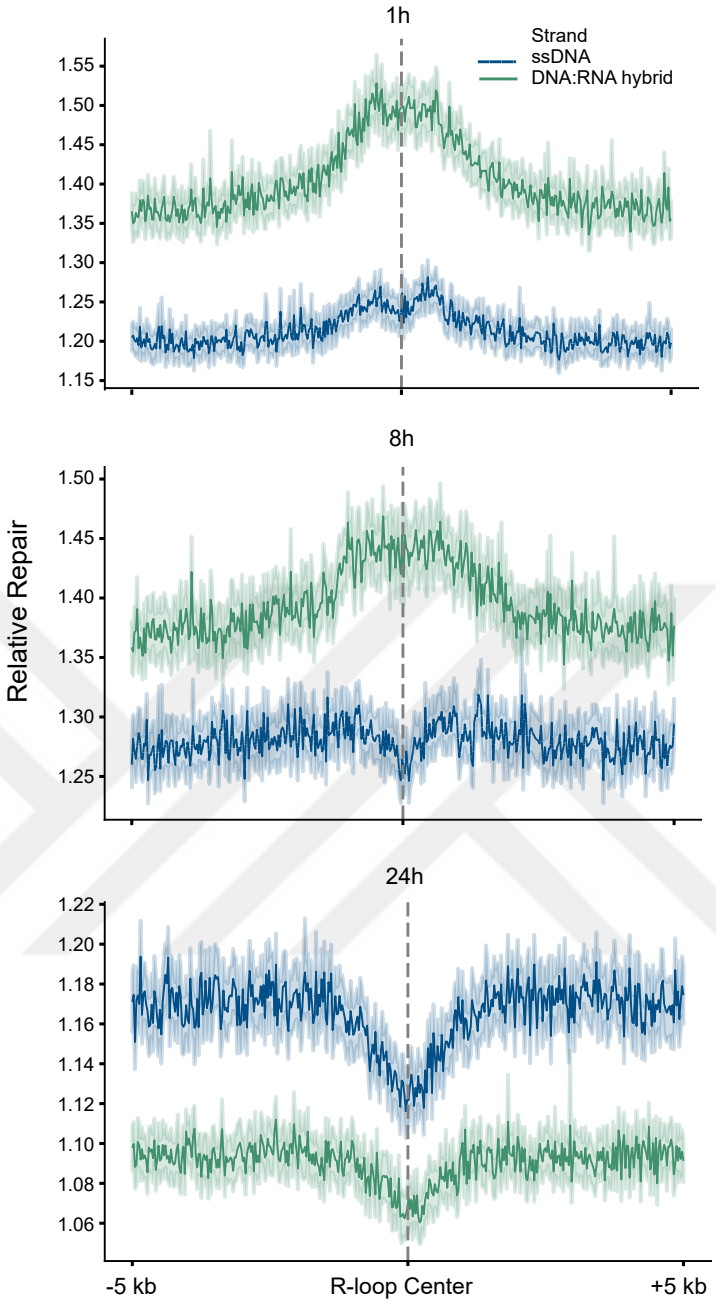

Figure 4.38 Repair profiles on RLBase R-loops in NHF1 cells. R-loop centers and 5 kb upstream and downstream regions were taken into consideration and divided into 400 equal bins. Damage-seq (1, 8, 24-hour time-points) and XR-seq (1, 8, 24-hour time-points) data reads were intersected with the genomic bins and the overlaps were counted and subjected to RPKM normalization. The overlapped Damage-seq and XR-seq RPKMs were first normalized by simulated data RPKMs to obtain normalized damage and normalized repair, respectively. Then, normalized repair was divided by normalized damage on each bin to obtain relative repair.

Since the effect of TC-NER was clear on the relative repair profiles on R-loops, eliminating the TC-NER effect would provide us a clear vision on how R-loops

impacted NER. For that reason, we have classified protein-coding genes according to their expression levels using the RNA-seq data obtained in our lab from HeLa cells and compared the relative repair between R-loop-containing and R-loop-lacking gene segments within the same expression subclass using the NHF1 1-hour time-point data (Figure 4.39, 'Wild-type' panel). By doing this, we aimed to equalize the TC-NER levels in the comparison since the gene segments with similar expression levels would receive similar TC-NER activities.

Another approach to eliminate the TC-NER effect from the comparison was the usage of an XR-seq data obtained from CSB knock-out NHF1 cells (Hu et al., 2017) at 1-hour time point. CSB is an essential protein for TC-NER pathway (Menoni, Wienholz, Theil, Janssens, Lans, Campalans, Radicella, Marteijn & Vermeulen, 2018). Its removal from the cell would lead to the inactivation of TC-NER. Therefore, we have calculated the relative repair rates on the gene segments that contain or lack R-loops, from the same expression subgroups using the CSB knock-out XR-seq data (Figure 4.39, 'CSB -/-' panel). Comparison of relative repair between gene segments containing and lacking R-loops revealed significant differences on both TS and NTS of gene expression subclasses TPM 5-10 and TPM < 1 where the gene segments containing R-loop were more efficiently repaired (Figure 4.39, 'Wild-type' panel). This result was expected since previous repair profiles revealed a higher relative repair on R-loop centers than on flanking sites (Figure 4.38), which altogether suggested that the could be a privilege on R-loop regions for repair.

The relative repair rate comparison with the CSB knock-out data revealed a significantly higher repair on the NTS of TPM 5-10 subclass genes whereas this difference was not observed on the TS of the same subclass (Figure 4.39, 'CSB -/-' panel). On the other hand, on the TPM < 1 subclass, relative repair on both TS and NTS showed a significantly higher repair on gene segments containing R-loops. These results showed that even when the TC-NER was inactivated, R-loop regions were repaired more efficiently. Therefore, the higher repair on R-loops could not be explained by higher TC-NER on those regions. Another evidence for this was that higher repair on the NTS of R-loop-containing gene segments being also observed. Since no TC-NER activity is expected on NTS, additional factors should be present to result in higher repair on those regions. The reason for the relative repair being at similar levels on the TS of TPM 5-10 subclass genes but being significantly different on the TS of TPM < 1 genes in CSB knock-out data might be an obstacle present on highly expressed genes preventing GG-NER, such as stalled RNA polymerases.

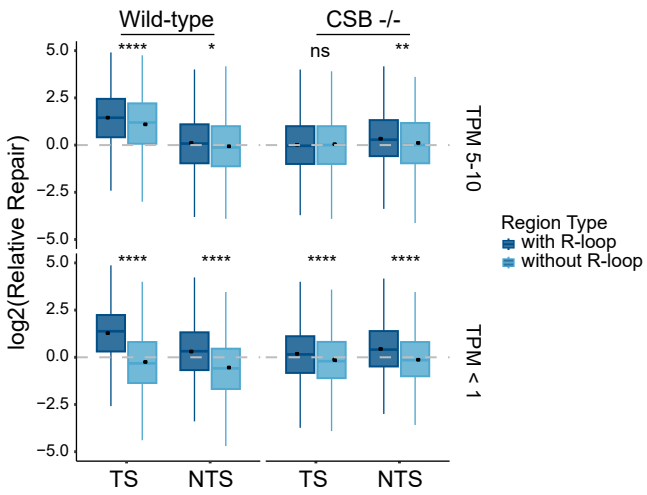

Figure 4.39 Relative repair differences between gene segments containing and lacking R-loops. Genes were overlapped with R-loop centers and spanning 2 kb regions and gene segments with overlaps were defines as 'with R-loops'. 'without R-loops' class of gene segments were randomly selected from genes with no overlap with any whole R-loop regions. Relative repair rates were calculated for each gene segment as explained in Figure 4.38. t-test was used to compute P-values. (ns: $p > 0.05$; *: $p <= 0.05$; **: $p <= 0.01$; ***: $p <= 0.001$; ****: $p <= 0.0001$)

The relative repair profiles on RLBase R-loops in CSB knock-out NHF1 cells showed the similar profile with HeLa (12-minute time-point) and wild-type NHF1 cells (1-hour time-point) (Figure 4.40, 4.37, 4.38). Unlike the profiles in HeLa and wild-type NHF1 cells, the strand difference in relative repair was little. ssDNA was repaired more efficiently than DNA:RNA hybrid on the flanking sites which might be due to the stalled RNA polymerases on TS blocking the repair process at some level. Since there would be no RNA polymerase stalling on NTS, the repair proteins would be reaching the damaged sites more efficiently. The repair being higher on R-loop centers even when the TC-NER was inactive further proved that the higher repair activity on R-loops was not solely due to TC-NER; there might be other factors playing roles in enhancing repair.

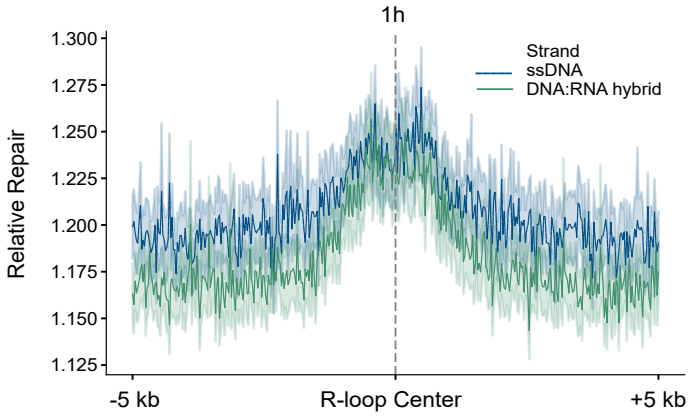

Figure 4.40 Relative repair profiles on RLBase R-loops in CSB knock-out NHF1 cells. R-loop centers and 5 kb upstream and downstream regions were taken into consideration and divided into 400 equal bins. Wild-type damage-seq (1-hour time-point) and CSB knock-out XR-seq (1-hour time-point) data reads were intersected with the genomic bins and the overlaps were counted and subjected to RPKM normalization. The overlapped Damage-seq and XR-seq RPKMs were first normalized by simulated data RPKMs to obtain normalized damage and normalized repair, respectively. Then, normalized repair was divided by normalized damage on each bin to obtain relative repair.

In summary, as we did with the damage profiles, we have compared the repair efficiencies on R-loops from different methods and databases and decided to continue the further analyzes with the RLBase R-loops since they are assessed by a set of quality-control criteria (Miller et al., 2022). On RLBase R-loops, we observed higher repair on R-loops when compared to their surrounding regions. In addition, DNA:RNA hybrid was repaired more efficiently than the ssDNA. The activation of TC-NER mechanism increased this difference in repair between the two strands. However, the repair difference between R-loops and their surroundings could not be explained by the higher TC-NER activity on those regions since the same difference was observed in CSB knock-out cells where TC-NER is inactive, indicating the presence of other factors enhancing the repair on R-loops.

4.2.3 Mutational burden on R-loops

Genomes of melanoma and other skin-related cancers suffer from a significant burden of single nucleotide mutations where the predominant type is cytosine-to-thymine (C>T) conversions. UV leads to lesions on DNA which is the main cause of these mutations when left unrepaired by NER machinery (Hayward, Wilmott, Waddell, Johansson, Field, Nones, Patch, Kakavand, Alexandrov, Burke & others, 2017; Hodis,

Watson, Kryukov, Arold, Imielinski, Theurillat, Nickerson, Auclair, Li, Place & others, 2012). We aimed to check the C>T mutations on RLBase R-loops since mutation accumulation would give idea about how effective NER mechanism had been on those regions. International Cancer Genome Consortium (ICGC) provides simple somatic mutations from the genomes of various cancer types obtained by whole-genome sequencing (WGS). From those, we have chosen three skin cancer cohorts (melanoma (MELA-AU), skin cutaneous melanoma (SKCM-US) and skin adenocarcinoma (SKCA-BR)) and two other non-skin cancer types (breast cancer (BRCA-US) and uterine corpus endometrial carcinoma (UCEC-US)). We have filtered the data to select C>T mutations since they were the targets of NER before turning into mutations. Since C>T mutation formation depends on the cytosine (C) presence, we have normalized the C>T mutation counts on each genomic region with the C counts on the same region.

In the C>T mutational profiles on R-loops, strand difference was observed in three skin-related cancers whereas no such difference was seen in BRCA-US and UCEC-US genomes (Figure 4.41). ssDNA strand contained more mutations than DNA:RNA hybrid strand in the genomes of skin-related cancers. This strand difference was also seen in relative repair profiles on NHF1 and HeLa cells where ssDNA had lower relative repair rate than DNA:RNA hybrid which is consistent with the mutational difference on the two strands (Figure 4.38, 4.37). MELA-AU and SKCA-BR genomes contained slightly lower C>T mutations on R-loop centers when compared to the flanking regions. On the other hand, SKCM-US, BRCA-US and UCEC-US contained higher mutations on R-loop centers than on flanking regions.

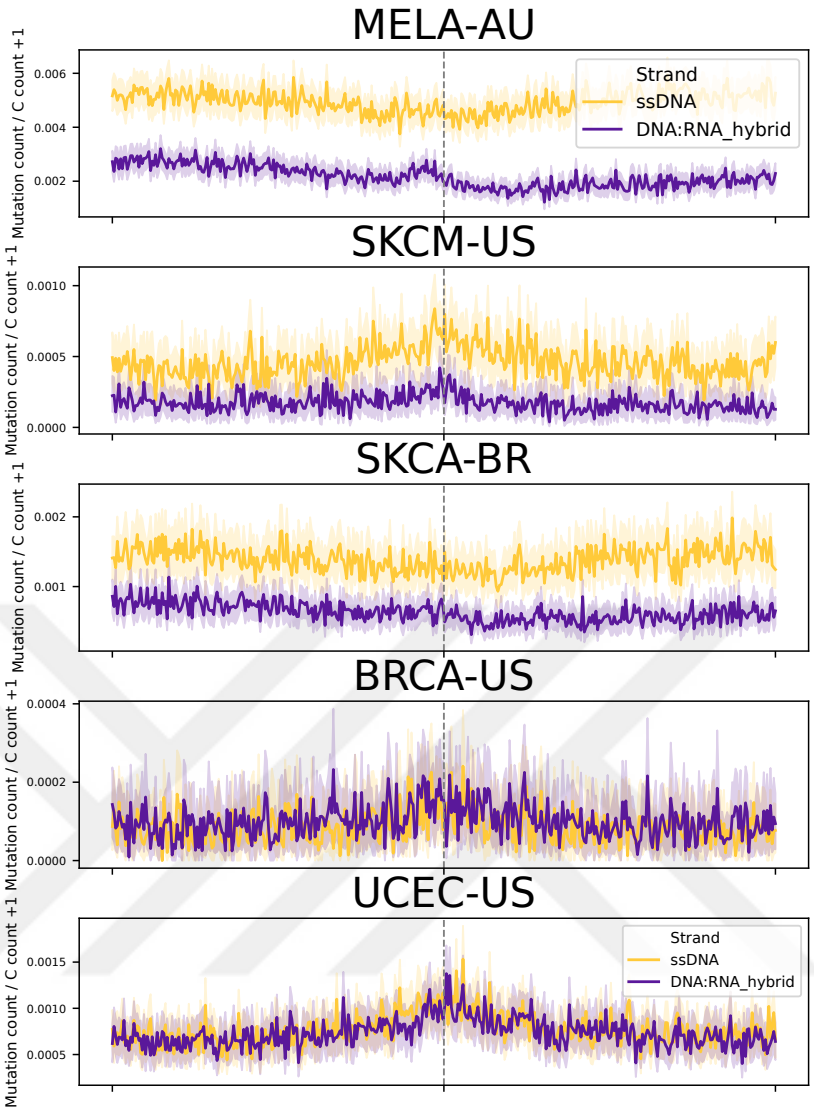

Figure 4.41 Mutational burden on R-loops. Distribution of C>T mutations on R-loops and -/+ 5 kb flanking regions in five cancer genomes: MELA-AU: melanoma, SKCM-US: skin cutaneous melanoma, SKCA-BR: skin adenocarcinoma, BRCA-US: breast cancer, UCEC-US: uterine corpus endometrial carcinoma. R-loop regions were divided into bins and intersected with C>T mutations on plus strand and G>A mutations on minus strand coming from each cancer data. Intersected mutation counts were normalized with cytosine nucleotide counts in each bin.

Higher NER activity and lower mutation accumulation could be due to the higher transcriptional activity on those regions which lead to a more active TC-NER. To figure out whether the lower mutation on R-loop centers was because of high TC-NER activity or some R-loop-related reason independent from the TC-NER, we needed to eliminate the TC-NER from the analysis. By using a set of genes with the same expression levels (TPM 5 to 10) and comparing the segments of those genes that contained R-loops or did not contain R-loops in terms of mutation burden,

we would see the direct effect of R-loops on repair and mutational accumulation (Figure 4.42). On both TS and NTS, R-loop-containing gene segments accumulated significantly less mutations than gene segments without R-loops in MELA-AU and SKCA-BR genomes. The same outcome was observed on the TS in SKCM-US cohort while on NTS, the difference was not significant. The genomes of non-skin-related cancers did not show significant differences between R-loop-containing genes and genes without R-loops, except for the TS in UCEC-US cohort. These results indicated that although differences in skin-related cancers was seen, this could be due to the types and subtypes of the cancers that were included in those cohorts. It was expected to see no significant differences in mutational accumulation in non-skin-related cancer since the C>T mutations in those genomes were probably not UV-related and not the targets for NER.

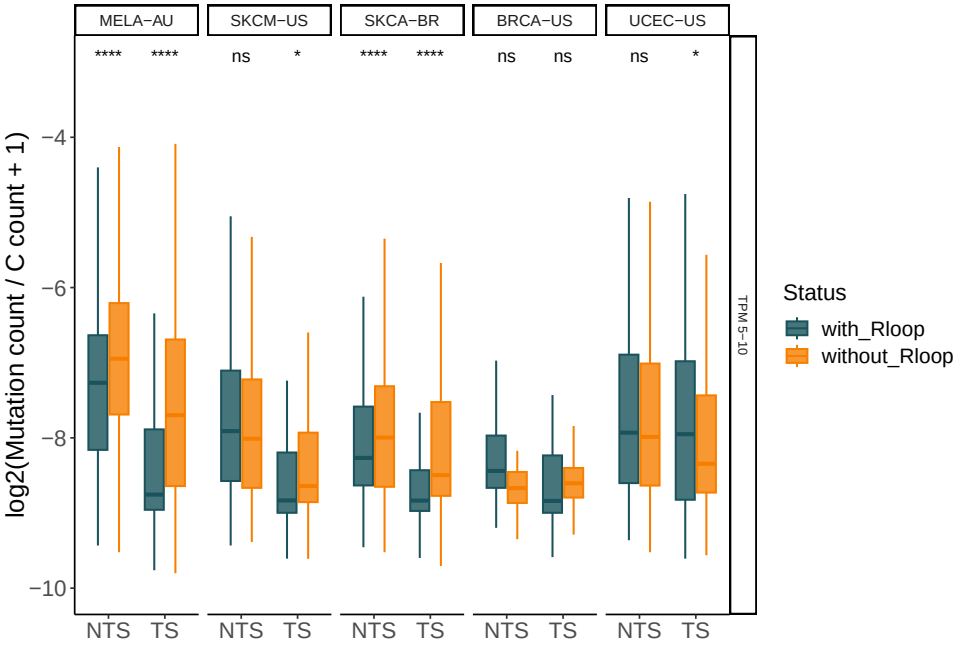

Figure 4.42 Mutational abundance on gene segments with/without R-loops. Genes were overlapped with R-loops and gene segments that overlap with an R-loop DNA:RNA hybrid on their TS were classified as 'genes with R-loops'. Genes that did not intersect with any R-loops were classified as 'genes without R-loops'. C>T mutations on plus strand and G>A mutations on minus strand in five cancer genomes (MELA-AU: melanoma, SKCM-US: skin cutaneous melanoma, SKCA-BR: skin adenocarcinoma, BRCA-US: breast cancer, UCEC-US: uterine corpus endometrial carcinoma) were counted on each gene segment and normalized by the C count.

Expected number of mutations differ on each region due to the difference in sequence content. Therefore, we calculated the expected C>T mutation numbers on gene segments with and without R-loops to compare with the observed C>T mutation counts using an algorithm adopted from Frigola et al. (Frigola et al., 2017). On

the gene segments with R-loops, the expected mutation counts were almost 3-folds lower than the expected mutation counts on gene segments without R-loops (Figure 4.43a, b). This demonstrated that regions with R-loops might be forming less mutations due to their sequence content. On the other hand, when we compared the observed and expected mutations on gene segments with R-loops, there were less mutation formation than expected (Figure 4.43a), indicating that although the sequence content lowers the expected mutation numbers on regions with R-loops, an additional factor should be present that even lowers the rate of mutation formation. This factor might be either a direct effect of R-loop presence or an indirect effect dependent of additional elements.

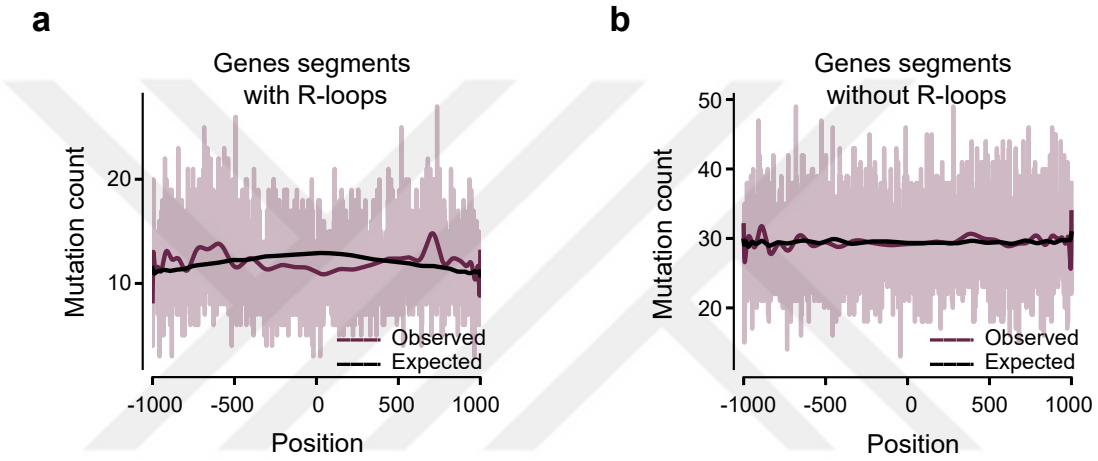

Figure 4.43 Observed and expected melanoma (MELA-AU) C-to-T mutation counts on gene segments. Observed and expected C->T mutation counts on gene segments that (a) contain or (b) lack R-loops were calculated using the algorithm adopted from Frigola et al. (Frigola et al., 2017).

In summary, R-loop regions accumulated less mutations than other regions and this accumulation was lower than expected by their sequence content. These observations were consistent with our previous findings that R-loop regions were repaired more efficiently than other regions, which could prevent UV-induced damages to turn into mutations frequently.

4.2.4 R-loop regulatory states on genome

Regulation of R-loops is a process which is involved by proteins from various pathways (Hegazy et al., 2020). The resolution of R-loops with the right timing is crucial

in order to avoid the transcription-replication conflicts and stalling of replication fork due to R-loop presence while benefiting from them in various molecular mechanisms (Hegazy et al., 2020). So far, we have gained an insight on how R-loop regions affect damage formation and NER activity. On the other hand, no study has been made to classify R-loops according to the proteins that regulate them. We aimed to test whether different subgroups of R-loops can be regulated by different combinations of regulator proteins. If this was possible, we wondered if these subgroups of R-loops were repaired with different efficiencies. To do this, we have gathered the R-loop regulators from the literature and retrieved the available ChIP-seq data as well as their control ChIP-seq data, if available (Table 3.1). We have processed the ChIP-seq reads and with the help of ChromHMM algorithm (Ernst & Kellis, 2017), we assigned genomic bins into states. To do that, we first tested different lengths of genomic bins (5 and 10 kb) and different number of states (5, 10 and 15) to create the model. By comparing the emmission and transition probabilities of these models, we decided that 5 kb-sized genomic bins and 10 states gave the best results where each state had distinctive characteristics for occupancy by different combinations of regulator proteins (Figure 4.44).

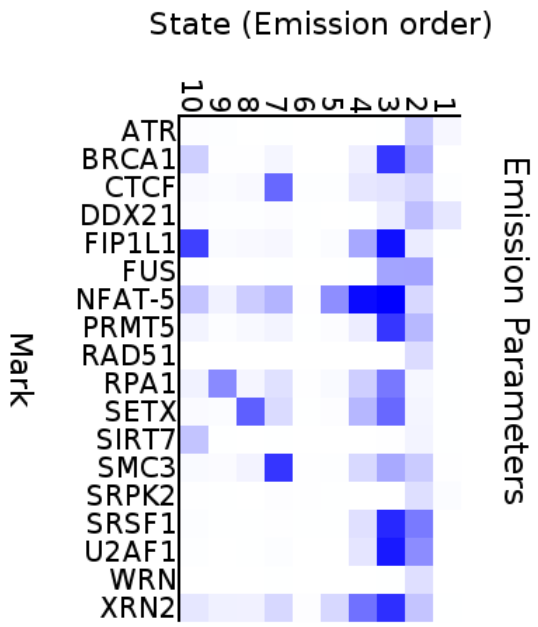

Figure 4.44 R-loop regulators included in the HMM model and the regulator occupancies on states. Darker colors represent higher protein occupancy and lighter colors represent lower protein occupancy.

In the HMM model, states 2, 3, 4 and 7 had the highest occupancies by different regulators while states 1 and 6 had the lowest protein occupancies (Figure 4.44).

The most dominant protein in state 7 were CTCF and SMC3 which are related to 3D genome organization (Rowley & Corces, 2018). Senataxin (SETX) binding was highly observed in state 8. SETX is a DNA:RNA helicase which has roles in transcription termination (Ramachandran, Ma, Griffin, Ng, Foskolou, Hwang, Victori, Cheng, Buffa, Leszczynska & others, 2021). In state 9, RPA subunit RPA1 binding was observed, which is a single-stranded DNA-binding protein that has roles in various DNA processes such as replication and recombination (Maréchal & Zou, 2015). RPA also binds the undamaged ssDNA during NER and positions XPA to the damage site (Topolska-Woś, Sugitani, Cordoba, Le Meur, Le Meur, Kim, Yeo, Rosenberg, Hammel, Schärer & others, 2020). In terms of genome coverage, state 6 had the highest length which covered almost 60% of the genome while state 1 had the second highest coverage (Figure 4.45).

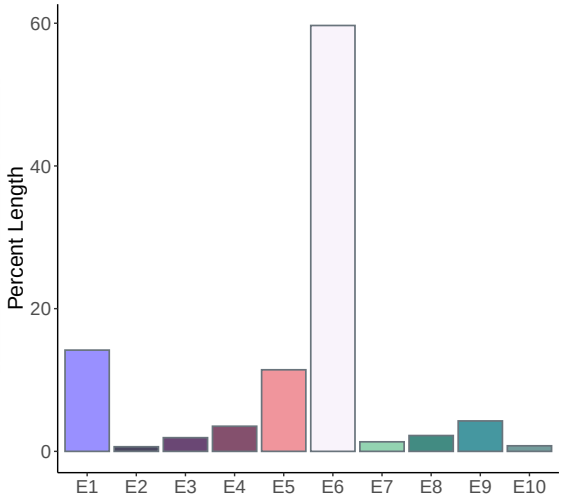

Figure 4.45 Percent genome lengths covered by the states.

The R-loop content of the states also differed. State 1 with the highest genome coverage and the lowest protein occupancy contained the highest portion of the R-loops (Figure 4.46). States 3, 4 and 5 also contained higher number of R-loops than other states. The lowest number of R-loops was located in state 1 which had the second lowest regulator occupancy.

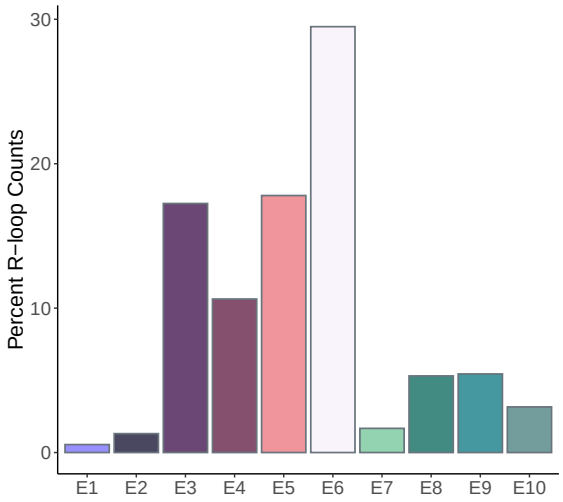

Figure 4.46 Distribution of R-loops into states. R-loops located on each state are counted and the percentage of total R-loops located on each state is calculated.

More than 10% of the state 3 segments contained R-loops which is the highest ratio among all states (Figure 4.47). On the other hand, ratio of segments that contained R-loops was the lowest in state 1 and state 6 which had the lowest regulator occupancy (Figure 4.44).

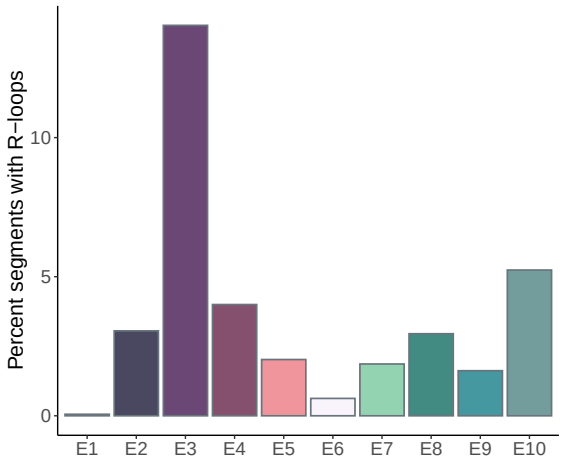

Figure 4.47 Portions of the states that contained R-loops. 5 kb bins of each state were checked for R-loop overlaps. The bins that overlap with at least 80% of the regions that included R-loop centers and 2 kb flanking sites were counted and percentages of R-loop-containing segments in each state was calculated.

To check how accessible the segments in the states were, we have used an ATAC-seq data obtained from HeLa cells (Li et al., 2021) and counted the intersections of the reads with the segments in states (Figure 4.48). After RPKM normalization, we have seen that the most accessible segments belonged to states 3, 4 and 7 which had

the highest protein occupancies. Segments in states 1, 2 and 6 segments were the least accessible according to the analysis. In general, the protein binding frequency (Figure 4.44) and the accessibility of the chromatin (Figure 4.48) correlated.

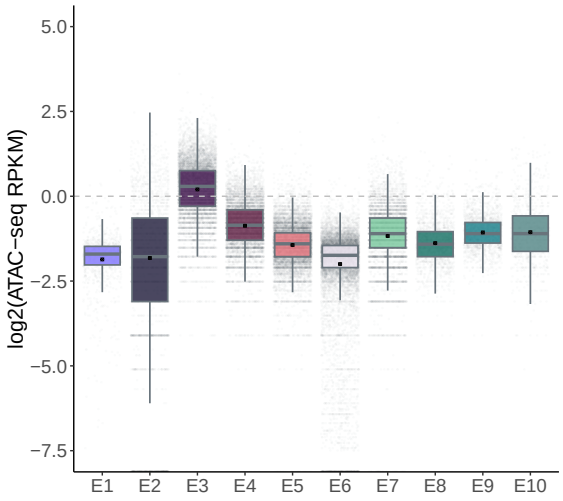

Figure 4.48 Chromatin accessibility levels of the states. ATAC-seq reads overlapping with each segment in the states were counted and RPKM normalization was performed for each count.

We also checked the accessibility levels of R-loops on each state relative to the non-R-loop regions on each state (Figure 4.49). The accessibility of R-loops on states followed a different pattern than the overall accessibility of the segments of states (Figure 4.48). Even though state 2 segments had lower accessibility, R-loops on the same state had the highest accessibility. R-loops on states 3 and 7 were also on more accessible chromatin than the R-loops on other states. Importantly, except for state 10 R-loops, R-loops from all states had higher accessibilities than the mean accessibility of their flanking regions on the same state.

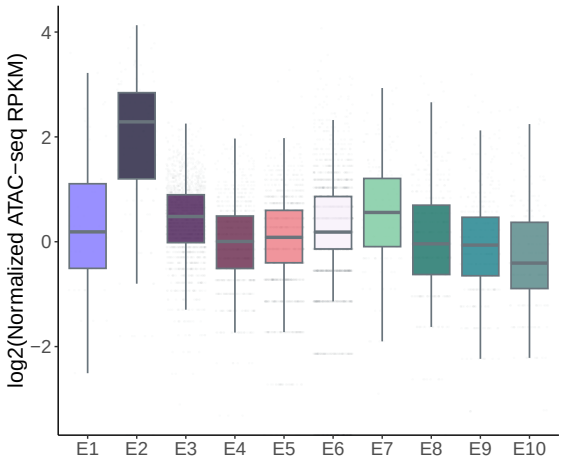

Figure 4.49 Chromatin accessibility levels of R-loops in the states. ATAC-seq reads overlapping with each R-loop in the states were counted and RPKM normalization was performed for each count. Random regions were selected from the non-R-loop-containing parts of the segments that contained R-loops. ATAC-seq reads were counted and RPKM normalization was performed. RPKMs on R-loops were normalized with the mean of RPKMs on non-R-loop-containing random parts in each state.

As demonstrated in Section 4.2.2., we have observed higher relative repair rates on R-loops when compared to their flanking regions. To check whether this difference was present in the states, we have calculated and compared the relative repair rates on R-loops and random regions from their flanking sites with both HeLa (Figure 4.50) and NHF1 data (Figure 4.51). Except for state 10 in HeLa, relative repair was significantly higher on R-loops than on random flanking regions in all of the states. This result was consistent with the relative chromatin accessibility results on the R-loop regions which showed higher relative accessibility on the R-loops from all states, except for the state 10 R-loops (Figure 4.49). On the other hand, the relative repair rates differed on R-loops in different states, indicating that even though R-loops in general followed the general profile, the relative repair rates were probably affected by the conditions variable between the states.

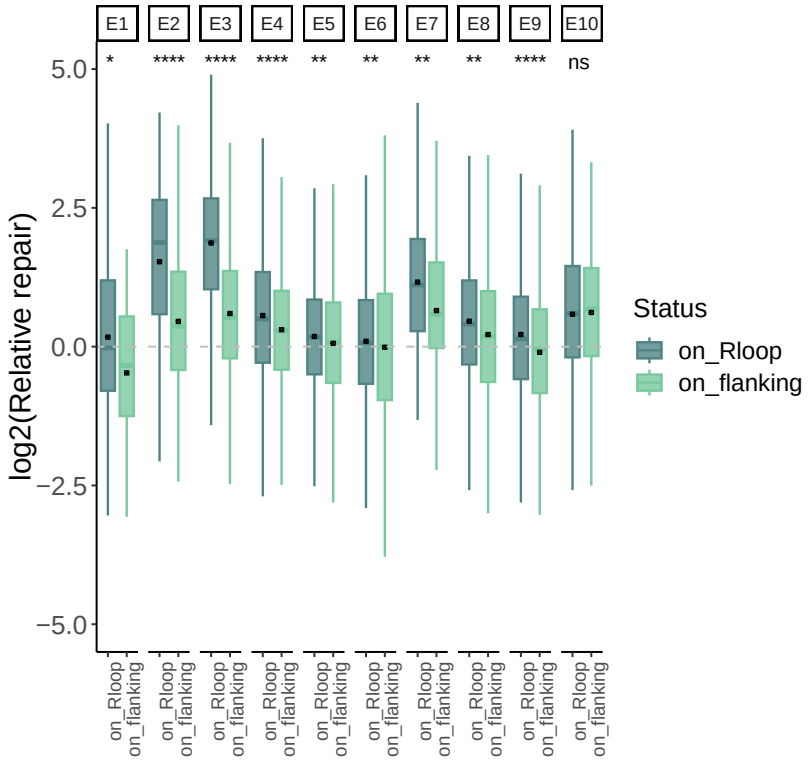

Figure 4.50 Comparison of the relative repair rates on R-loops and random non-R-loop-containing regions in states in HeLa cells. Relative repair rates were calculated by counting the Damage-seq (0-hour), XR-seq (12-minute) and simulated Damage-seq and XR-seq reads on each region. After RPKM normalizations, Damage-seq RPKMs were normalized by the simulated Damage-seq RPKMs to obtain normalized damage ratios for each region. Same thing was repeated with XR-seq and simulated XR-seq data to obtain normalized repair values. Finally, normalized repair on each region was divided by the normalized damage value to obtain relative repair rate. The significance between relative repair rates on R-loops ('on_Rloop') and random regions ('on_flanking') was computed with t-test (ns: $p > 0.05$; *: $p <= 0.05$; **: $p <= 0.01$; ***: $p <= 0.001$; ****: $p <= 0.0001$).

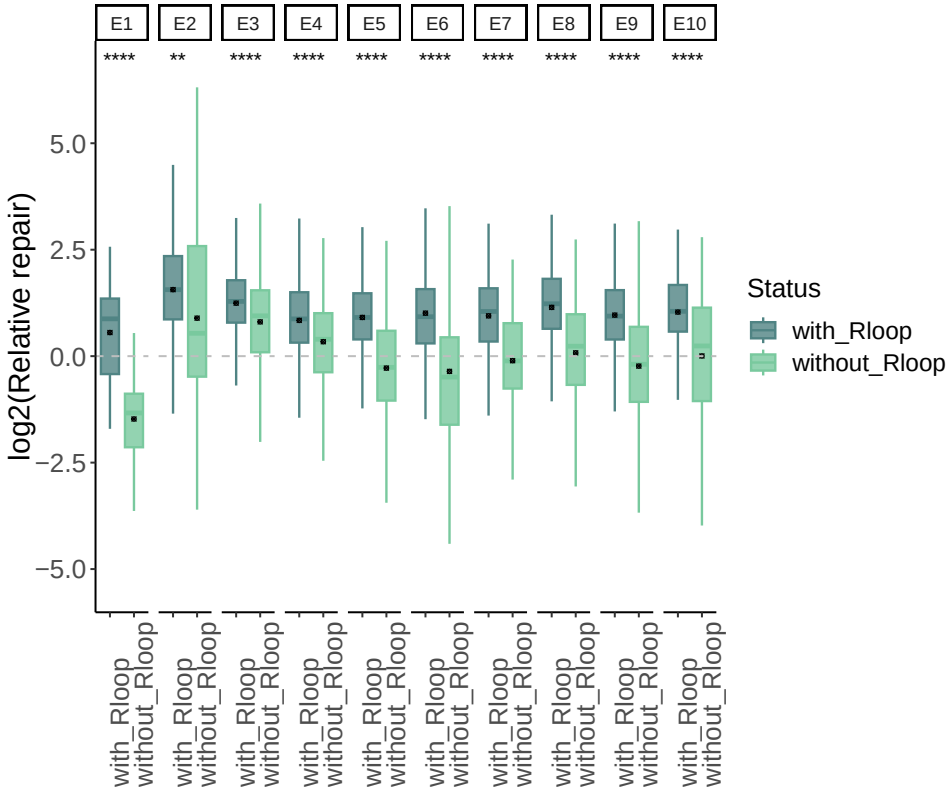

Figure 4.51 Comparison of the relative repair rates on R-loops and random non-R-loop-containing regions in states in NHF1 cells. Relative repair rates were calculated by counting the Damage-seq (1-hour), XR-seq (1-hour) and simulated Damage-seq and XR-seq reads on each region. After RPKM normalizations, Damage-seq RPKMs were normalized by the simulated Damage-seq RPKMs to obtain normalized damage ratios for each region. Same thing was repeated with XR-seq and simulated XR-seq data to obtain normalized repair values. Finally, normalized repair on each region was divided by the normalized damage value to obtain relative repair rate. The significance between relative repair rates on R-loops ('on_Rloop') and random regions ('on_flanking') was computed with t-test (ns: $p > 0.05$; *: $p <= 0.05$; **: $p <= 0.01$; ***: $p <= 0.001$; ****: $p <= 0.0001$).

To see how the NER efficiency differed between the R-loops of different states, we have normalized the relative repair rate on each R-loop with the mean of relative repair rates of random non-R-loop regions on the flanking sites of R-loops and compared them (Figure 4.52, 4.53). The analysis with HeLa (12-minute time-point) and NHF1 (1-hour time-point) revealed different results. The highest two relative repair were observed in state 2 and 3 R-loops while the lowest two were on state 6 and 10 R-loops in HeLa cells (Figure 4.52). In NHF1 cells, the highest two relative repairs were observed on state 1 and 3 R-loops while the lowest relative repair rates were on state 2, 6 and 10 R-loops (Figure 4.53). Except for states 2, 3 and 6, the relative repair on R-loops increased relative to other states between HeLa 12-minute and NHF1 1-hour time-points, which could be due to the activation of TC-NER on

those regions. TC-NER activities might be at various levels on different states due to differential features between the states. A decrease in relative repair was observed on state 2 and 3 R-loops where this decrease is more pronounced on state 2. State 2 R-loops were the most accessible one among R-loops in all states (Figure 4.49); therefore, the repair might be finalized on state 2 R-loops before the 1-hour time point which led us to observe the loss of repair efficiency at that time-point. In addition, repair activity on state 6 R-loops remained very similar between the two time-points. Since there were no regulator binding on this state (Figure 4.44), it can be speculated that state 6 R-loops were in a genomic region where the number of protein-coding genes were very low. Hence, the TC-NER mechanism might not be active there while GG-NER activity remained stable between the two time-points (Figure 4.52, 4.53).

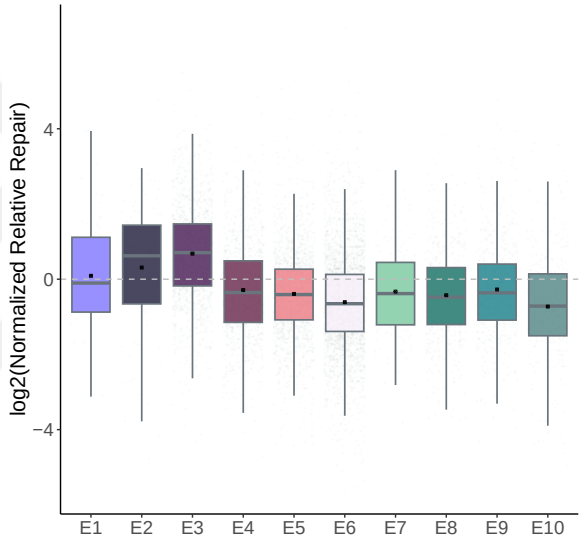

Figure 4.52 Comparison of the relative repair rates on R-loops among states in HeLa cells. Relative repair rates were calculated by counting the Damage-seq (0-hour), XR-seq (12-minute) and simulated Damage-seq and XR-seq reads on each R-loop. After RPKM normalizations, Damage-seq RPKMs were normalized by the simulated Damage-seq RPKMs to obtain normalized damage ratios for each region. Same thing was repeated with XR-seq and simulated XR-seq data to obtain normalized repair values. Finally, normalized repair on each region was divided by the normalized damage value to obtain relative repair rate. The same calculations were repeated for random non-R-loop regions chosen from the flanking sites of R-loops. The relative repair of each R-loop was divided by the mean of relative repair rates of random flanking sites on the same state. The significance between relative repair rates on R-loops among states was computed with t-test (ns: $p > 0.05$; *: $p <= 0.05$; **: $p <= 0.01$; ***: $p <= 0.001$; ****: $p <= 0.0001$).

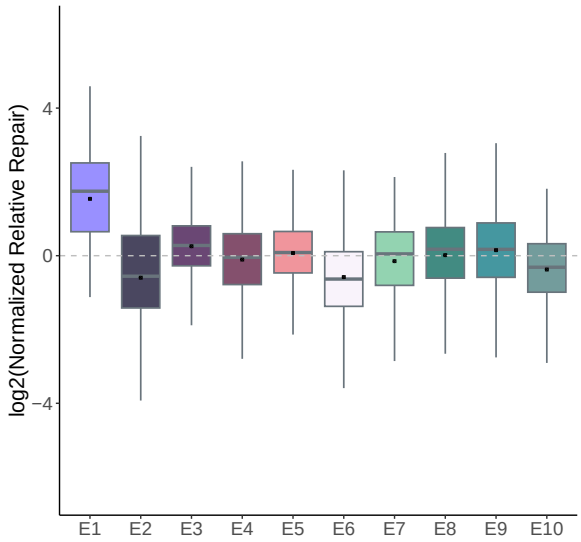

Figure 4.53 Comparison of the relative repair rates on R-loops among states in NHF1 cells. Relative repair rates were calculated by counting the Damage-seq (1-hour), XR-seq (1-hour) and simulated Damage-seq and XR-seq reads on each R-loop. After RPKM normalizations, Damage-seq RPKMs were normalized by the simulated Damage-seq RPKMs to obtain normalized damage ratios for each region. Same thing was repeated with XR-seq and simulated XR-seq data to obtain normalized repair values. Finally, normalized repair on each region was divided by the normalized damage value to obtain relative repair rate. The same calculations were repeated for random non-R-loop regions chosen from the flanking sites of R-loops. The relative repair of each R-loop was divided by the mean of relative repair rates of random flanking sites on the same state. The significance between relative repair rates on R-loops among states was computed with t-test (ns: $p > 0.05$; *: $p <= 0.05$; **: $p <= 0.01$; ***: $p <= 0.001$; ****: $p <= 0.0001$).

The relative repair rates displayed an increasing profile on R-loop centers at both 12-minute (HeLa) and 1-hour (NHF1) time-points as mentioned in Section 4.2.2 (Figure 4.37, 4.38). In addition, strand difference was observed where DNA:RNA hybrid strand and its flanking sites received higher repair efficiency than ssDNA and its flanking sites. This strand difference was more pronounced at 1-hour time-point due to the TC-NER activation on TS on genes while a smaller strand difference was observed at 12-minute time-point even though the TC-NER might not be fully active (Figure 4.37, 4.38). Consistently, on the R-loops of all regulatory states, relative repair on R-loop strands showed a higher differentiation in NHF1 1-hour time-point (Figure 4.54b) than in HeLa 12-minuste time-point (Figure 4.54a). However, the increased repair efficiency on R-loop centers compared to the flanking sites was not at the same intensity in every state. At both time-points, the peaks with the highest intensities were observed on the R-loops on states 2, 3 and 7 (Figure 4.54a, b) on which the R-loops with the highest accessibilities were observed (Figure 4.49). Furthermore, high regulator binding was observed on those states (Figure 4.44).

Interestingly, the strand difference on those states were less than the strand difference on other states. Altogether, these findings suggested that the R-loops which were located on more open chromatin might be the reason for the general trend on R-loops as being more efficiently repaired. Being on more open chromatin might also explain the lower strand difference on state 2, 3 and 7 R-loops, providing a better chance for NER proteins for accessing the damaged DNA on both strands. In addition, a slight decrease on top the peaks at the R-loop centers was seen in the general relative repair profiles (Figure 4.37, 4.38). This decrease was only seen on state 3 R-loops (Figure 4.54a, b) which had the highest regulator occupancy (Figure 4.44).

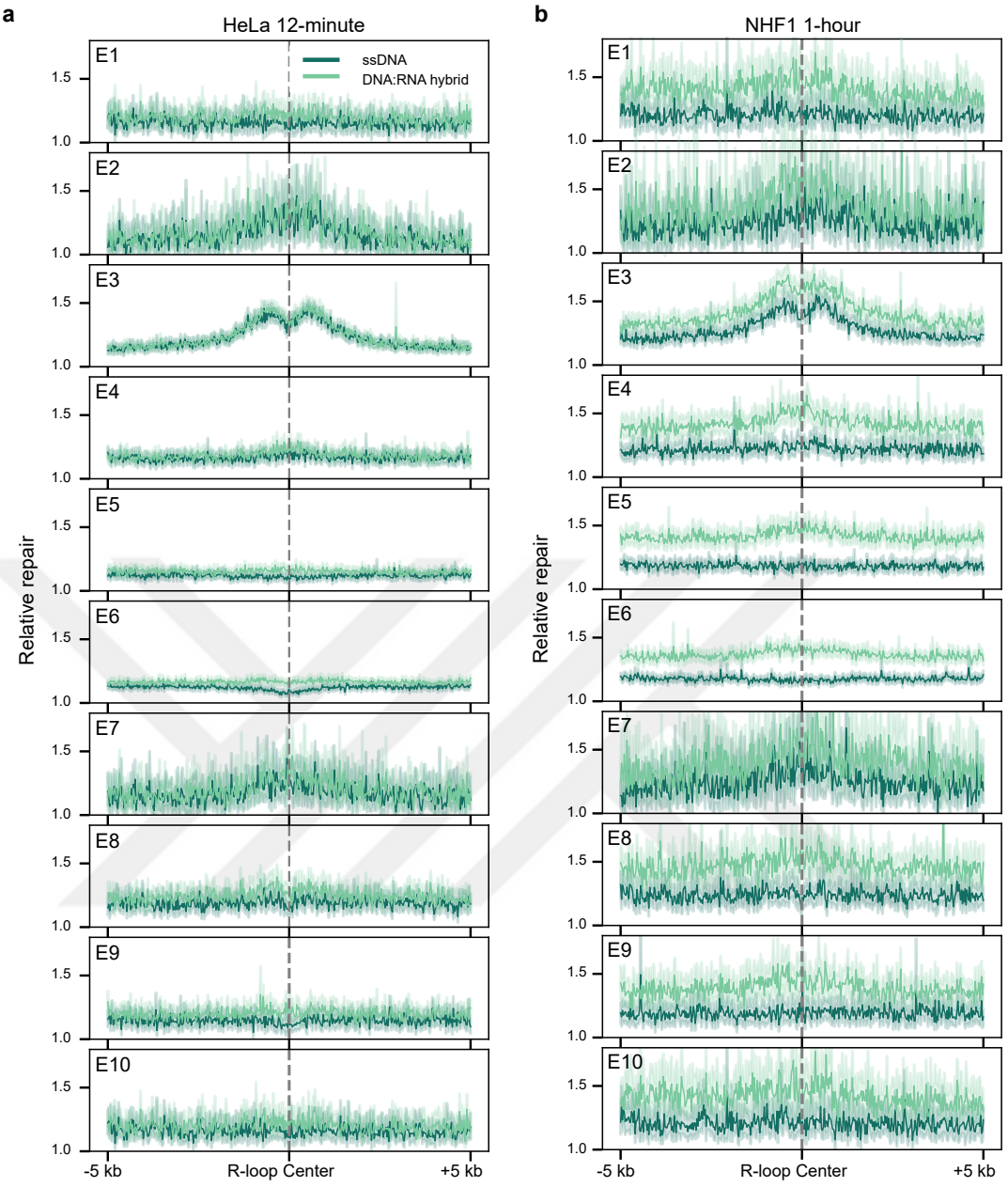

Figure 4.54 Repair profiles on R-loops from the states. R-loop centers and 1 kb flanking sites were intersected with the segments of the states and assigned to states if at least 80% of the region length including R-loop centers and 2 kb flanking sites was completely on a single state. R-loop centers and 5 kb upstream and downstream regions were taken into consideration and divided into 400 equal bins. Damage-seq and XR-seq data reads were intersected with the genomic bins and the overlaps were counted and subjected to RPKM normalization. The overlapped Damage-seq and XR-seq RPKMs were first normalized by simulated data RPKMs to obtain normalized damage and normalized repair, respectively. Then, normalized repair was divided by normalized damage on each bin to obtain relative repair. Colors represent the relative repair on the two R-loop strands (a) Relative repair profiles using XR-seq (12-minute) and Damage-seq (0-hour) data obtained from HeLa cells. (b) Relative repair profiles using XR-seq (1-hour) and Damage-seq (1-hour) data obtained from NHF1 cells.

When TC-NER was inactive, the same relative repair peaks were detected on R-loop centers of states 2, 3, 4 and 7 without a strand difference as expected (Figure 4.55). This supported that TC-NER was not the only reason why we observed enhanced repair activity on R-loop centers. However, not observing the same enhancement on every state led us doubt that the reason was not solely the presence of R-loops. The consistence between high ATAC-seq levels and high relative repair rates on the R-loops of state 2, 3 and 7 also suggested that there could be additional factors increasing the efficiency of repair in those three states. Oppositely, the relative repair peaks on the R-loops of other six states were lost when TC-NER was inactivated (Figure 4.55), indicating that the higher relative repair efficiency on the R-loops of different states might have different reasons. While the R-loops with higher chromatin accessibility retained the repair efficiency due to the GG-NER activity, R-loops with low accessibility lost the repair efficiency when the TC-NER is unavailable.

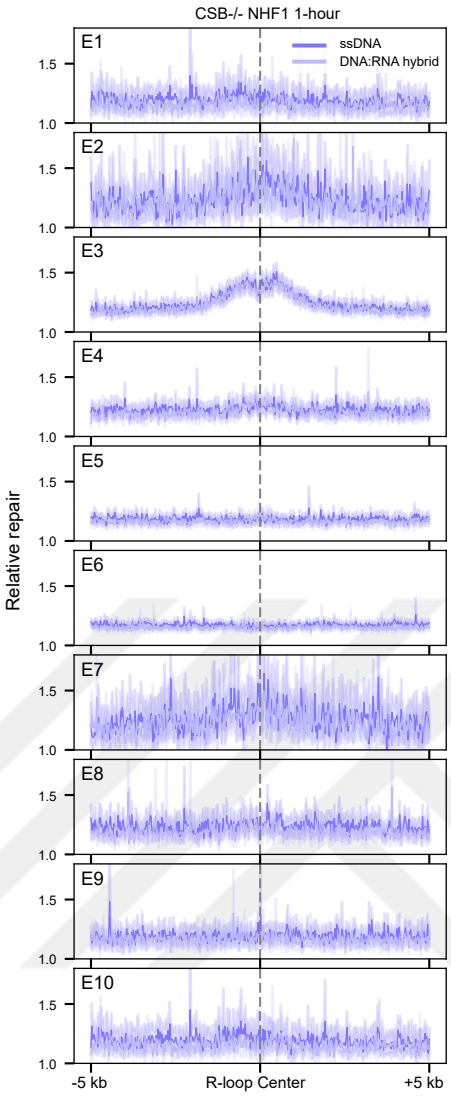

Figure 4.55 CSB knock-out repair profiles on R-loops from the states. R-loop centers and 1 kb flanking sites were intersected with the segments of the states and assigned to states if at least 80% of the region length including R-loop centers and 2 kb flanking sites was completely on a single state. R-loop centers and 5 kb upstream and downstream regions were taken into consideration and divided into 400 equal bins. Damage-seq (1-hour, wild-type NHF1) and XR-seq (1-hour, CSB knock-out NHF1) data reads were intersected with the genomic bins and the overlaps were counted and subjected to RPKM normalization. The overlapped Damage-seq and XR-seq RPKMs were first normalized by simulated data RPKMs to obtain normalized damage and normalized repair, respectively. Then, normalized repair was divided by normalized damage on each bin to obtain relative repair. Colors represent the relative repair on the two R-loop strands.

We have also compared the damage profiles on R-loop strands located on the states. As discussed in part 4.1.1. we have observed a difference in damage abundance on ssDNA, DNA:RNA hybrid and dsDNA structures (Figure 4.25, 4.26). In both HeLa and NHF1 0-hour time-points, ssDNA contained the lowest level of damage and

DNA:RNA hybrid contained the highest. On the R-loops from most of the states, the general damage distribution profile was observed (Figure 4.56a, b). However, with both HeLa and NHF1 0-hour time-point data, strand difference on R-loops from states 2, 3 and 4 was lower than the difference on other states. In addition, the damage abundance was lower on the R-loops from these three states than their flanking regions. States 2, 3 and 4 were the states with high regulator occupancies (Figure 4.44); therefore, this pattern could be due to frequent protein binding events.

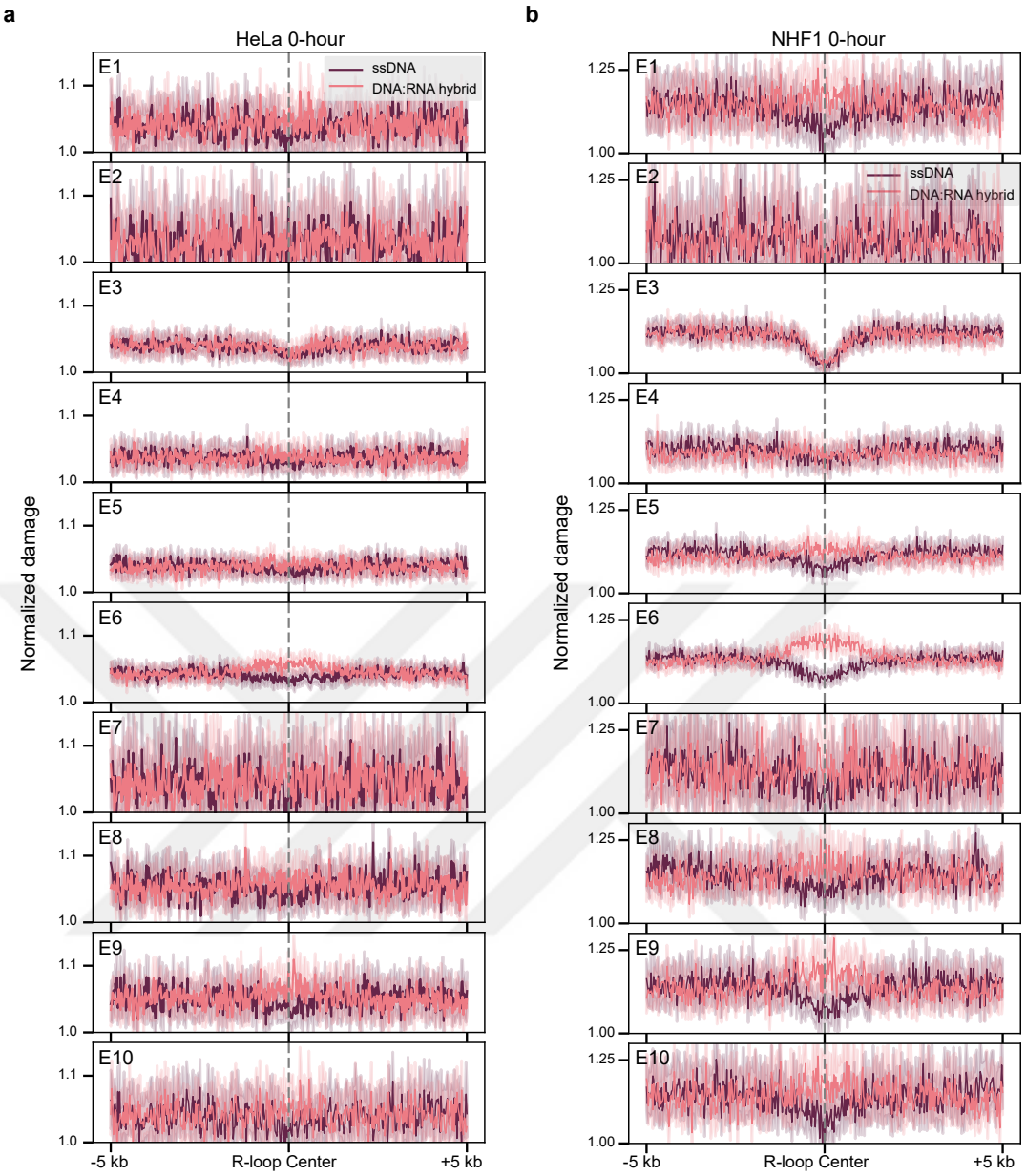

Figure 4.56 Damage profiles on R-loops from the states. R-loop centers and 1 kb flanking sites were intersected with the segments of the states and assigned to states if at least 80% of the region length including R-loop centers and 2 kb flanking sites was completely on a single state. R-loop centers and 5 kb upstream and downstream regions were taken into consideration and divided into 400 equal bins. Damage-seq data reads were intersected with the genomic bins and the overlaps were counted and subjected to RPKM normalization. The overlapped Damage-seq RPKMs were normalized by simulated data RPKMs to obtain normalized damage. Colors represent the normalized damage on the two R-loop strands. (a) Normalized damage profiles using Damage-seq (0-hour) data obtained from HeLa cells. (b) Normalized damage profiles using Damage-seq (0-hour) data obtained from NHF1 cells.

To sum up, the analysis of R-loops that were regulated by different combinations of

regulators revealed a correlation between relative repair and chromatin accessibility profiles. On the other hand, the relative repair seemed not affected from regulator combinations on the states. Notably, the intensity of the regulator occupation on R-loops correlated with the slight decrease in relative repair as well as the low damage formation on both R-loop strands.

4.3 Distribution of R-loops on *Arabidopsis* genome

Plant genomes also prevalently form R-loops which play important roles in various cellular pathways as human R-loops. Plant R-loops were identified in various processes such as genome integrity, gene expression, embryogenesis and 3D genome organisation (Cheng, Wang, Yao & Sun, 2021; Sun et al., 2013; Zhou, Lei, Shafiq, Zhang, Li, Li, Zhu, Dong, He & Sun, 2023). As we did for the human R-loops, we retrieved *Arabidopsis* R-loop positions from the literature, that were sequenced with the ssDRIP-seq method (Xu et al., 2017) which was one of the two available data at that time. We processed the raw reads and performed peak calling. We aimed to analyze the distribution of R-loops on different chromatin regions of the *Arabidopsis* genome and to assess the dataset in terms of the previously established features of R-loops.

Firstly, we checked the distribution of the R-loop sites on different chromatin features. We retrieved the locations of the nine chromatin states (Sequeira-Mendes et al., 2014) from the literature and counted the R-loop sites on those states. We also normalized the counts by the total lengths of the states since larger states may contain higher number of R-loops. The results showed that R-loops accumulated mostly on TSS, promoter and genic regions (Figure 4.57). On the contrary, the lowest number of R-loops were found on heterochromatin and intergenic regions as expected by the nature of R-loop structures (Castillo-Guzman & Chédin, 2021).

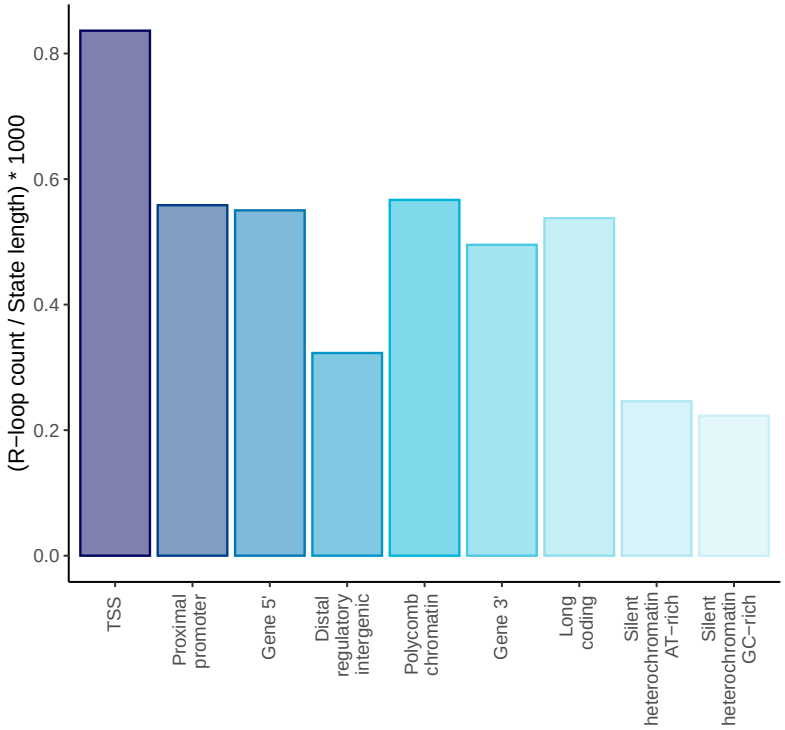

Figure 4.57 R-loop contents of *Arabidopsis* chromatin states. Intersections between R-loops and chromatin states were counted if at least 80% of the R-loop was located on the state. The counts on each state were divided by the total state length and multiplied by 1000 to obtain the normalized counts.

Secondly, we assessed transcriptional activity around R-loops using a GRO-seq dataset (Zhu et al., 2018). Since most of the R-loops form around regions where transcription is active (Hegazy et al., 2020), GRO-seq signal is expected to be high around R-loops. We observed high GRO-seq signal on both strands of the R-loops (Figure 4.58). Since the most R-loops form during transcription, DNA:RNA hybrid strands are usually located on the TS where the nascent RNA anneals. Therefore, we expected to see higher GRO-seq signal on the ssDNA since the GRO-seq reads should align on NTS during genome alignment. However, DNA:RNA hybrid contained a comparable signal (Figure 4.58a). We speculated that the reason for this could be the antisense transcription events due to the frequent intersection of genes on opposite strands. This is more expected on *Arabidopsis* genome because of the more compact genome of *Arabidopsis* when compared to the human genome. On the other hand, GRO-seq signal being higher on R-loops than their surroundings showed that these set of R-loops were located on transcriptionally active sites, as expected.

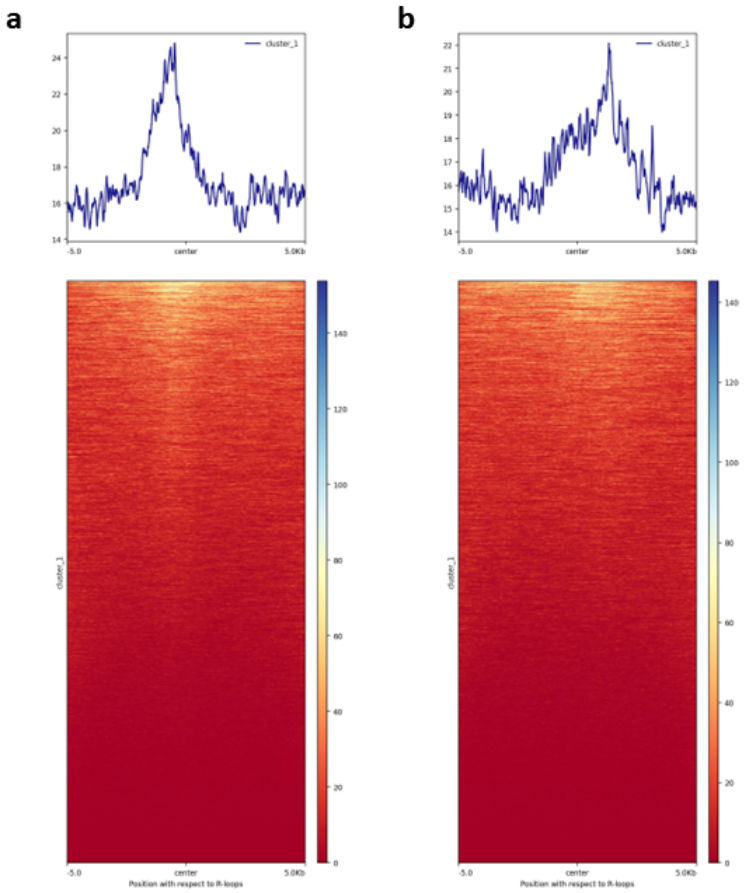

Figure 4.58 Transcription activity on *Arabidopsis* R-loops. Intersections between R-loops and GRO-seq data reads were counted and plotted using deepTools (Ramírez et al., 2014). (a) GRO-seq profiles on ssDNA strand. (b) GRO-seq profile on DNA:RNA hybrid strand.

However, from the heatmaps, we observed that this high signal pattern was not true for most of the R-loops (Figure 4.58a, b; lower panels). To check this, we clustered the R-loops according to their GRO-seq signal and showed that only a small subset of R-loops had very high or mild transcription signal while other R-loops had very low or no transcription on them (Figure 4.59a, b). This was observed on both strands of the R-loops.

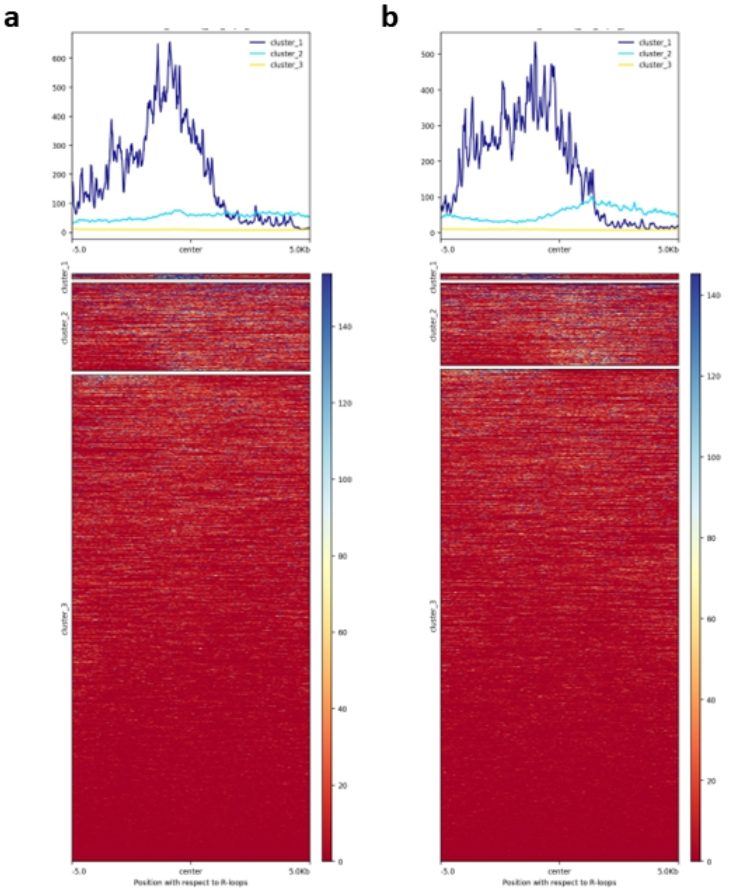

Figure 4.59 Transcription activity on clustered *Arabidopsis* R-loops. Intersections between R-loops and GRO-seq data reads were counted and plotted using deepTools (Ramírez et al., 2014). (a) GRO-seq profiles on ssDNA strand. (b) GRO-seq profile on DNA:RNA hybrid strand.

ssDRIP-seq R-loops satisfied the general criteria for the distribution of R-loops on the genome. They were mostly located on genic regions although a subset of R-loops were located on heterochromatin and intergenic regions (Figure 4.57). On the other hand, even though most of the R-loops were expected to be located on the regions where the transcription was active, only some of them satisfied these criteria (Figure 4.58). This might indicate that the data might be containing trans-R-loops which form on a different region from where the RNA was produced. It might also indicate some accuracy problems with the data possibly due to the low specificity of S9.6 antibody. However, we decided to use this set of R-loops for further analysis and assess them further in the context of UV-induced damage formation and repair in the proceeding sections.

4.4 Repair profiles in *Arabidopsis*

Plants encounter high levels of UV light as they spend most of their days under the sun. They also use nucleotide excision repair (NER) to cope with the damage caused by UV (Oztas et al., 2018). The mechanism of NER is well-characterized in mammalian cells while plant NER mechanism is yet to be studied. In order to understand more about how the *Arabidopsis* genome is protected from the harmful effects of UV-induced damages, we have analyzed NER mechanism in *Arabidopsis* (Kaya et al., 2022).

Due to the identified role of CSA protein in transcription-coupled nucleotide excision repair (TC-NER) in human cells, CSA1 and CSA2 might potentially involve in the TC-NER in *Arabidopsis* (Zhang, Guo, Zhang, Guo, Schumaker & Guo, 2010). To shed light onto the roles of CSA1 and CSA2 proteins in TC-NER, we conducted a comparative analysis of sequence similarities between CSA1 and CSA2, as well as their similarities with other proteins found in eukaryotic organisms. To achieve this objective, we retrieved the most similar eukaryotic protein sequences for both the CSA1 and CSA2 proteins. We conducted the sequence alignments (Figure 4.60) and constructed a phylogenetic tree (Figure 4.61). Multiple sequence alignment identified 33 distinct locations where CSA1 and CSA2 differed (Figure 4.60). These positions had low conservation scores when compared to other positions in the alignment. The residues found in sites with high conservation scores were predominantly common among CSA1 and CSA2 protein sequences, indicating that these two proteins are not markedly distinct from each other. The existence of selection pressure against conserved sites of CSA-family in the CSA2 protein indicated that these two related proteins probably maintain the same or similar functions, potentially in distinct tissues. In addition, the phylogenetic tree revealed that the CSA gene duplication occurred recently in *Arabidopsis* as the CSA1 and CSA2 was clustered together in the *Arabidopsis* clade (Figure 4.61).

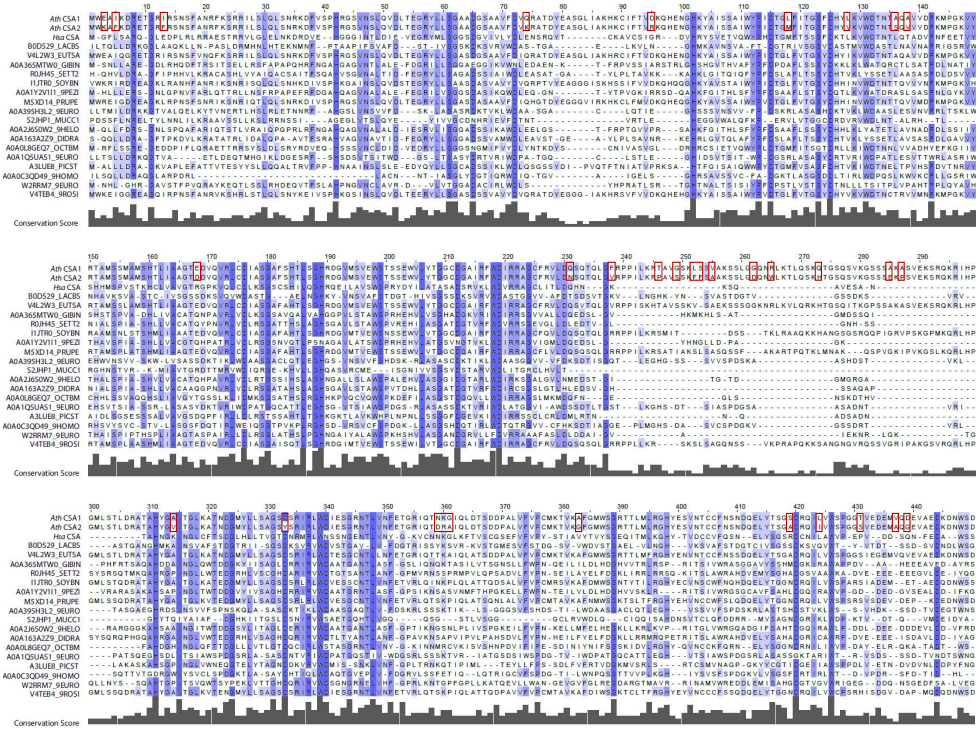

Figure 4.60 Multiple sequence alignment for *Arabidopsis* CSA1 and CSA2 and their corresponding proteins. The *Arabidopsis* CSA1 and CSA2 proteins are aligned with the 1000 most similar eukaryotic protein sequences using MAFFT. Conservation scores are then determined using Bio3D. For visual purposes, a set randomly selected 17 proteins, *Arabidopsis* CSA1 and CSA2, as well as human CSA, were displayed together. Furthermore, the image does not incorporate the gap sites seen in the *Arabidopsis* CSA1 and CSA2 genes. Dark hues represent residues that are strongly conserved, whereas light hues suggest reduced conservation across all protein sequences in the multiple sequence alignment. The discrepancies in the protein sequences of *Arabidopsis* CSA1 and CSA2 are highlighted by the red boxes.

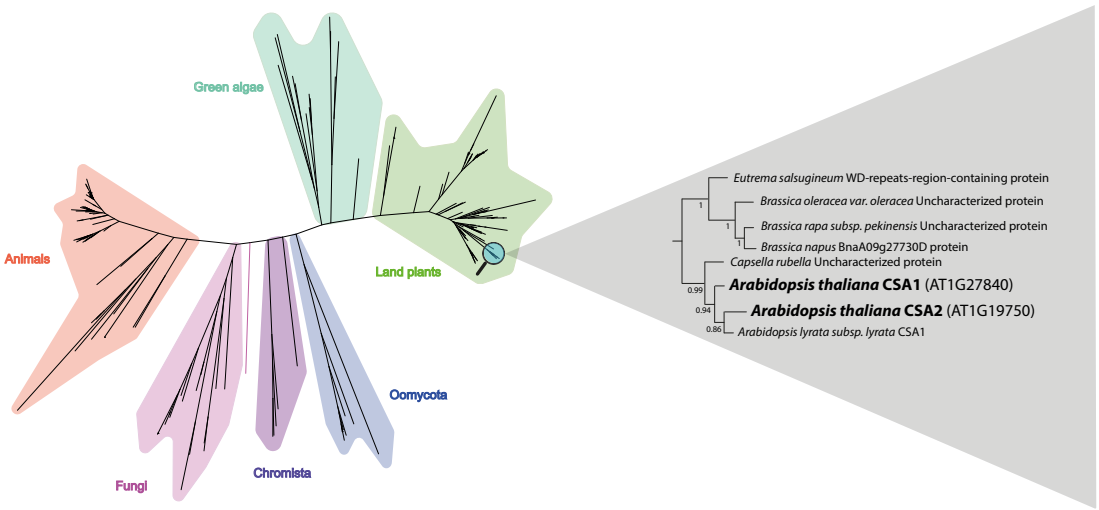

Figure 4.61 Conservation of CSA proteins in eukaryotes. A phylogenetic tree is constructed using IQ-TREE2 with 1000 bootstrap replicates (-B 1000). The tree includes 200 eukaryotic protein sequences that exhibited the highest similarity to *Arabidopsis* CSA1 and CSA2 proteins. The *Arabidopsis* CSA1 and CSA2 proteins, together with their most closely related orthologs, are displayed in a distinct tree, which includes the corresponding bootstrap values.

4.4.1 Repair activity on *Arabidopsis* genes

Understanding how the genic regions are maintained in *Arabidopsis* is important to gain more insight about the NER mechanism in plants. To do that, we obtained XR-seq datasets for genome-wide CPD damage from wild-type *Arabidopsis* plants 1 hour after the UV exposure. In addition, to shed light onto the players of NER pathways, we obtained *csa1* and *csa2* knock-out *Arabidopsis* plants as well as double-mutants. By comparing the XR-seq data from mutant and wild-type plants, we gained more idea about the importance of CSA1 and CSA2 proteins in UV-induced CPD damage repair by NER pathway.

Firstly, we checked the quality of the XR-seq datasets by assessing their nucleotide contents and read length distributions (Figure 4.62). We observed an increase in the concentration of "TT" at the 4-6 nucleotide position from the 3' end of the oligonucleotides, as anticipated (Canturk, Karaman, Selby, Kemp, Kulaksiz-Erkmen, Hu, Li, Lindsey-Boltz & Sancar, 2016). The size distribution of excision products in each sample was comparable with the earlier study (Oztas et al., 2018), with lengths ranging from 23 to 27 nucleotides. Additionally, a population of DNA fragments with lengths between 14 and 20 was found, which is likely due to the degradation of the excision products. The dinucleotide content and size distribution of excision prod-

ucts indicated that the dual incision process in *csa1* and *csa2* is the same as in the wild-type.

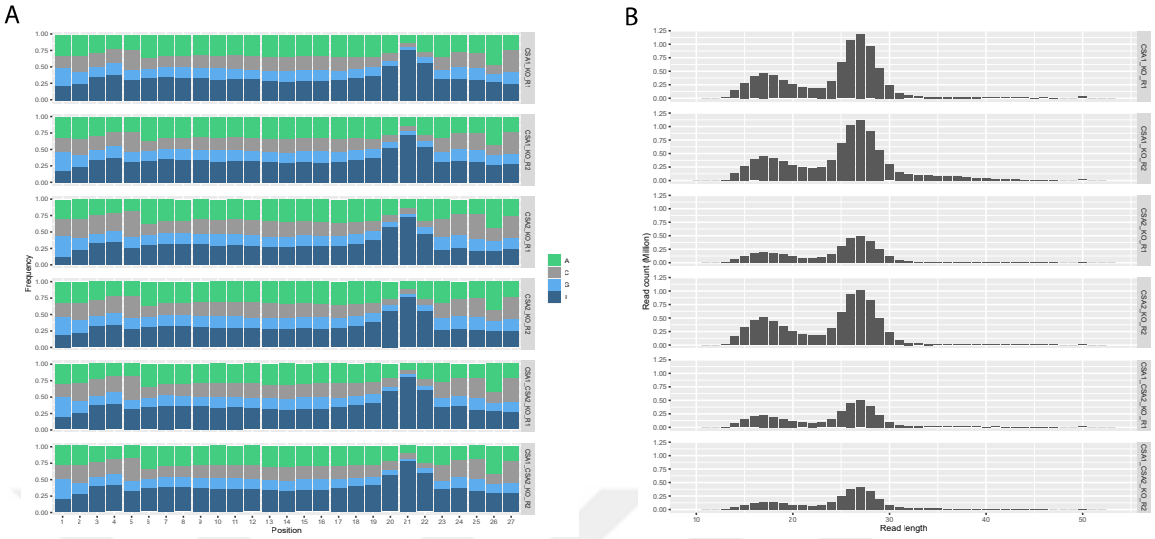

Figure 4.62 Nucleotide content and length distribution in XR-seq datasets. (A) Nucleotide frequencies of 27 nucleotide-long XR-seq reads coming from *csa1*, *csa2*, and *csa1csa2* plants. (B) Length distribution of XR-seq reads for *csa1*, *csa2*, and *csa1csa2* XR-seq data.

To test the dependence of TC-NER on CSA1 and CSA2 proteins, we compared the XR-seq profiles of the mutant and wild-type plants (Figure 4.63). Calculations of TS/NTS repair on *EIF4G* gene revealed a substantial decline in TC-NER signal in *csa1* plants, whereas TC-NER levels exhibited a minor but significant reduction in *csa2* plants compared to the wild type (Figure 4.63a). In addition, the TS/NTS repair in the double-mutant samples were in comparable levels with the *csa1* sample.

Similar differences were also observed between the mutant and wild-type samples when all protein-coding genes were included in the analysis (Figure 4.63b). The highest TC-NER level was observed in wild-type plants whereas the TC-NER level in *csa2* mutant plants slightly decreased. The repair in *csa1* and *csa1csa2* mutants were inhibited when compared to wild-type and *csa2* mutant samples whereas the repair in *csa1csa2* mutant was even lower than the repair in *csa1* mutant. These differences between samples were also observed when the TS/NTS repair ratios were calculated for the protein-coding genes, which were also statistically significant (Figure 4.63c).

TC-NER levels can also be affected by the sequence content of the genomic region of interest (Canturk et al., 2016). To eliminate this possibility, simulated XR-seq data of wild-type and *csa1* samples, which were created by taking into account the nucleotide content of the real XR-seq data and randomly selecting reads from

the genome (Akkose & Adebali, 2023), were also included in the analysis (Figure 4.63b). The levels of simulated XR-seq data were very low on protein-coding genes, indicating that the higher repair levels were not due to the sequence content.

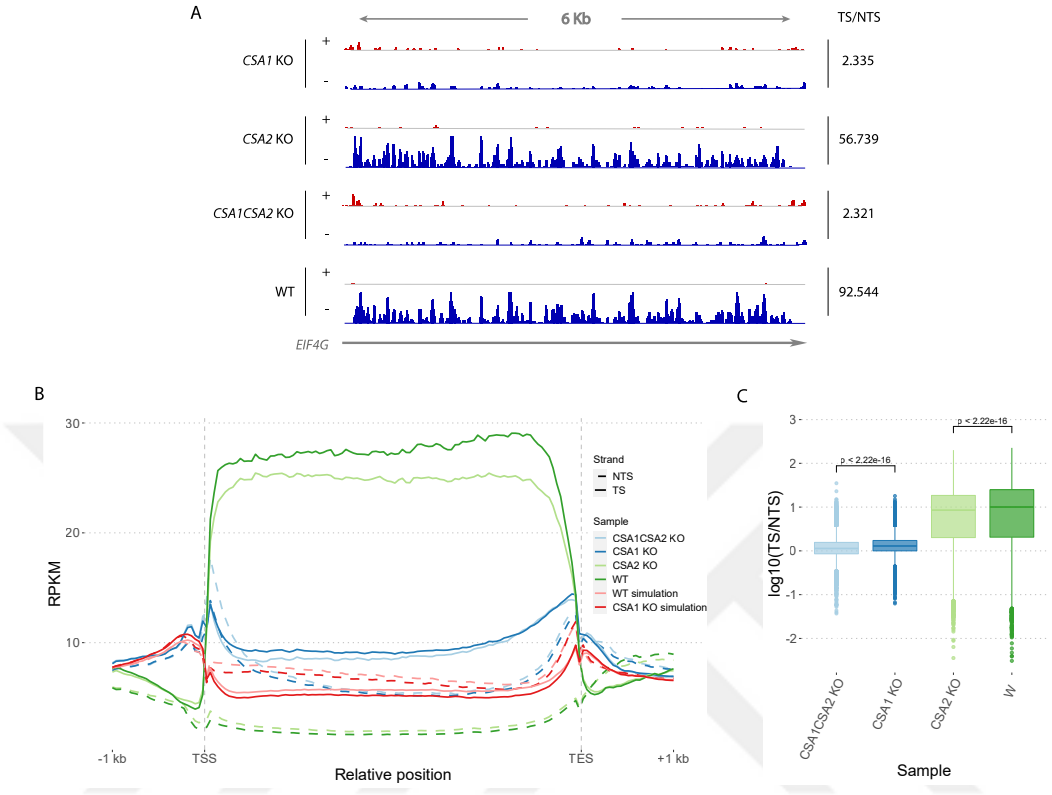

Figure 4.63 Repair profiles on *Arabidopsis* genes. (A) Image capturing repair profiles of the *EIF4G* house-keeping gene. The Integrative Genomics Viewer is employed to visually represent strand-specific XR-seq data from wild-type and *csa1*, *csa2*, and *csa1csa2* knock-out samples. The representation of plus and minus DNA strands is denoted by the symbols + and –, respectively. The *EIF4G* gene is shown by a gray horizontal line, and the arrow denotes its direction on the genome. The TS/NTS ratios for each sample on the *EIF4G* gene are provided on the right side. (B) XR-seq profiles on protein-coding genes in the Araport11 genome. The read counts were subjected to RPKM normalization. The genic regions encompass both 1 kb upstream of the transcription start site (TSS) and downstream of the transcription end site (TES). The dataset includes XR-seq data for both wild-type and knock-out samples, as well as simulated XR-seq data for simulated *csa1* and wild-type samples. The two replicates from each sample are combined. (C) TS/NTS repair ratios on Araport11 protein-coding genes. The merged replicates of knock-out and wild-type XR-seq samples have been included. The significance of the difference between samples is assessed using the Wilcoxon test. RPKM: Reads Per Kilobase of transcript per Million mapped reads; TSS, Transcription start site; TES; Transcription end site.

The TC-NER levels are directly correlated to the transcription levels on the genes (Oztas et al., 2018). Thus, the higher repair on *csa2* mutant plants could be due to

higher transcription activities in those plants when compared to *csa1* mutant plants. To eliminate this doubt, we checked the TC-NER levels on csa1 and csa2 genes in mutant and wild-type samples (Figure 4.64). In wild-type sample, the repair was significantly reduced on csa2 gene which could be due to the lower expression of this gene.

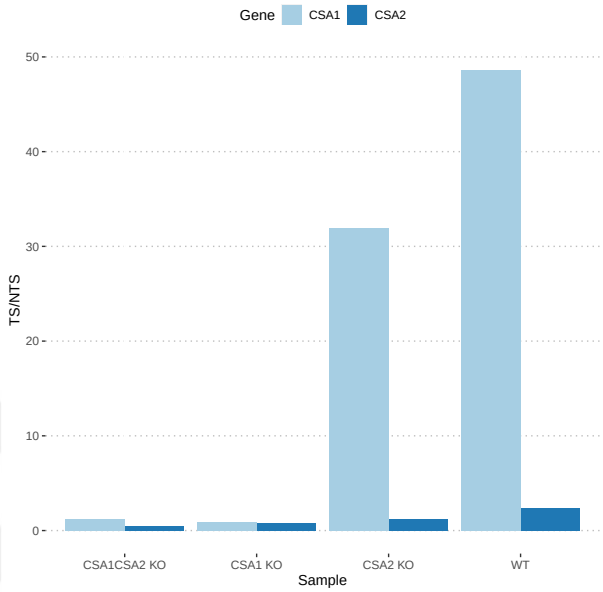

Figure 4.64 Transcription levels of *csa1* and *csa2* genes based on XR-seq data. The ratio of the repair levels on transcribed strand (TS) to non-transcribed strand (NTS) on csa1 and csa2 genes in wild-type, *csa1*, *csa2* and *csa1csa2* mutant plants.

In summary, our analysis showed that CSA1 is required for an effective TC-NER, whereas CSA2 has a minimal impact on TC-NER. We observed a little reduction in TC-NER levels in *csa2* plants compared to wild-type, suggesting that the CSA2 protein is not essential for TC-NER function, despite the previous evidence of its interaction with CSA1 (Zhang et al., 2010). The similarity in TC-NER level difference between wild-type and *csa2*, as well as between *csa1* and *csa1csa2* samples, indicated the genuine decrease in TC-NER in the absence of CSA2.

4.4.2 Repair activity on *Arabidopsis* R-loops

In the previous sections, we examined the ssDRIP-seq data that provided the R-loop position on *Arabidopsis* genome. We also analyzed the TC-NER activity on *Arabidopsis* genes and the roles of CSA1 and CSA2 in TC-NER. After those analysis, we aimed to test the impact of R-loops on NER in *Arabidopsis*, as we did with human

R-loops previously. For that purpose, we used the XR-seq data of wild-type and *csa1* and *csa1csa2* mutant *Arabidopsis* plants and checked the repair profiles on R-loops. We did not include *csa2* mutant XR-seq data in the analysis due to the discrepancies between the two replicates. We performed normalizations with the simulated XR-seq data and obtained normalized repair which eliminated the potential sequence content bias. However, we were not able to normalize the repair by the damage abundance since no Damage-seq data was available for *Arabidopsis*.

The normalized repair profiles on *Arabidopsis* ssDRIP-seq R-loops showed differentiation between wild-type and mutant samples as well as between the two R-loop strands (Figure 4.65). The repair was higher in the wild-type samples than the repair in mutants, which was expected due to the inhibition of TC-NER in mutant samples. Notably, the repair levels of wild-type and mutant samples did not merge at the 2 kb flanking regions of R-loop centers, which might be due to the high abundance of genes on those regions causing them to be repaired mainly by TC-NER. This observation was also expected as we knew that R-loops are the products of transcription (Hegazy et al., 2020) and in addition, as we observed in our previous analysis that the distribution of R-loops were highly abundant on genic regions (Figure 4.57) where the transcription was active (Figure 4.58). Furthermore, normalized repair levels were also different between the *csa1* and *csa1csa2* mutants. As we have discussed in our previous study, CSA proteins play an important role in the recruitment of TC-NER factors to the damaged site (Kaya et al., 2022). Both CSA1 and CSA2 losses negatively affect TC-NER, while at different levels (Figure 4.63) (Kaya et al., 2022). The mutant repair profiles on R-loops were consistent with the repair profiles on genes. A reduced normalized repair was observed on R-loops in *csa1* mutant whereas in *csa1csa2* mutant, the normalized repair levels were even lower, supporting the importance of CSA1 and CSA2 in the repair of R-loop regions. Moreover, the reduction of repair on R-loops suggested that R-loops are mostly repaired by TC-NER.

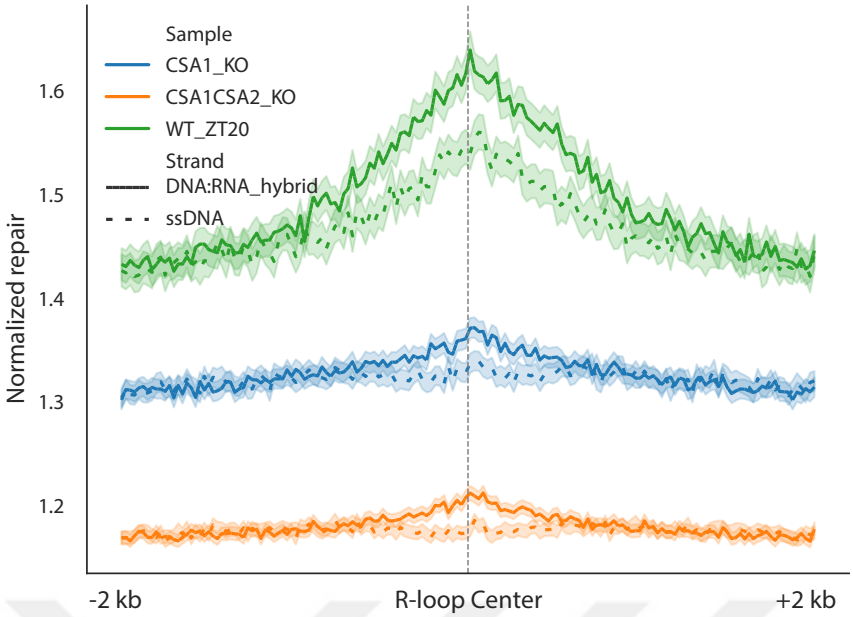

Figure 4.65 Repair profiles on *Arabidopsis* R-loops. Normalized repair calculations were performed using wild-type ('WT_ZT20'), *csa1* ('CSA1_KO') and *csa1csa2* ('CSA1CSA2_KO') mutant XR-seq data and the simulated XR-seq data for each sample. The ssDRIP-seq peak centers and 2 kb upstream and downstream regions were included in the analysis. The regions were divided into 160 equal bins and the intersections with the XR-seq reads were checked. The intersecting reads were counted strand-specifically and subjected to RPKM normalization. Normalized repair was obtained by dividing the real XR-seq RPKMs by the simulated XR-seq RPKMs on each bin and each strand, separately.

Apart from the normalized repair level differentiation between the wild-type and mutant plants, a repair peak was observed at the R-loop centers and immediate adjacent regions in all conditions (Figure 4.65). These peaks were consistent with the peaks observed on human RLBase R-loops (Figure 4.37, 4.38). In addition, the length distribution of *Arabidopsis* R-loops revealed that most of the R-loops lie within the regions spanning 1 kb around the centers (Figure 4.66). Thus, the normalized repair peaks were likely present the effect of R-loops as both peaks coincided. In addition, the peak in wild-type normalized repair did not fully disappear in the mutant normalized repair profiles as we observed in the CSB knock-out human NHF1 cells (Figure 4.40), suggesting that other factors, such as chromatin accessibility, might promote repair on R-loop centers as discussed in Section 4.2.4. In addition, as observed in human RLBase R-loops, DNA:RNA hybrid strand had a higher normalized repair than the ssDNA, while this difference disappeared on the distal flanking regions (Figure 4.65). This result can normally be explained by the presence of TS of genes, which are repaired by TC-NER, on the DNA:RNA hybrid strand commonly, as the R-loop hybrids form when the nascent RNA anneals on

the template strand. However, we detected gene abundance on both strands of the *Arabidopsis* R-loops (Figure 4.67), indicating that the strand difference in repair levels might be due to some other reason. Notably, the lack of knowledge about the damage levels on these regions prevent us to assess repair levels accurately. In general, the repair on *Arabidopsis* R-loops resembled the relative repair rate profiles on human RLBase R-loops in terms of intensity increase at the centers and the strand difference.

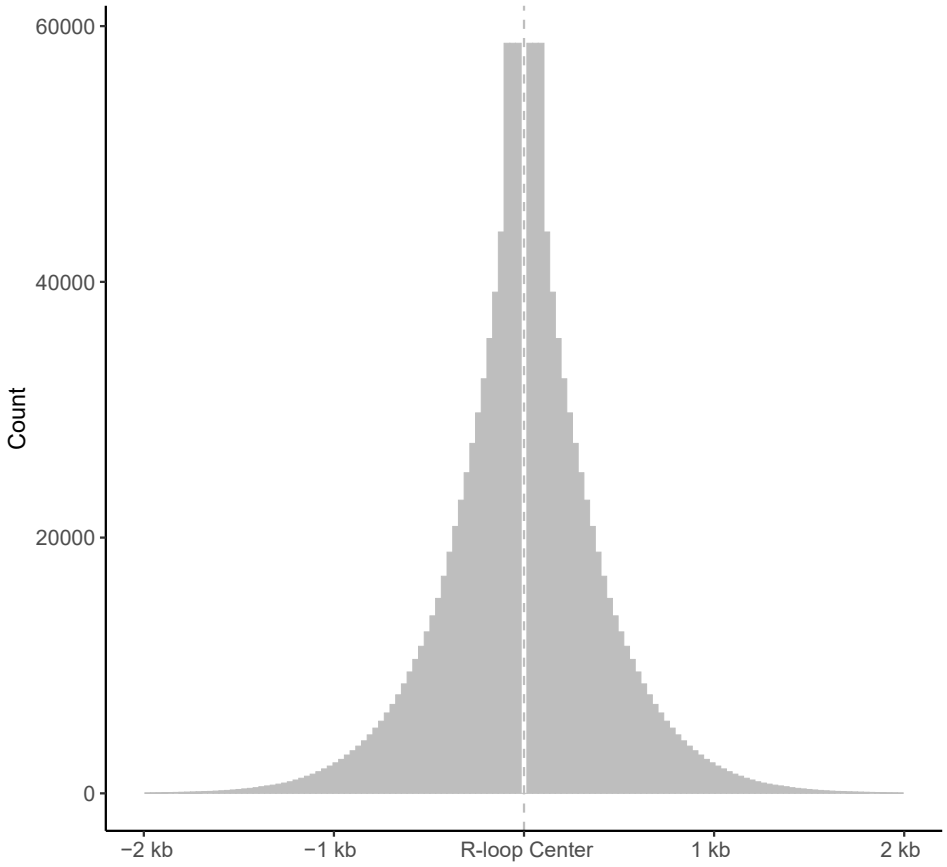

Figure 4.66 Length distribution of *Arabidopsis* R-loops. The regions around R-loop centers covered by the peaks were plotted.

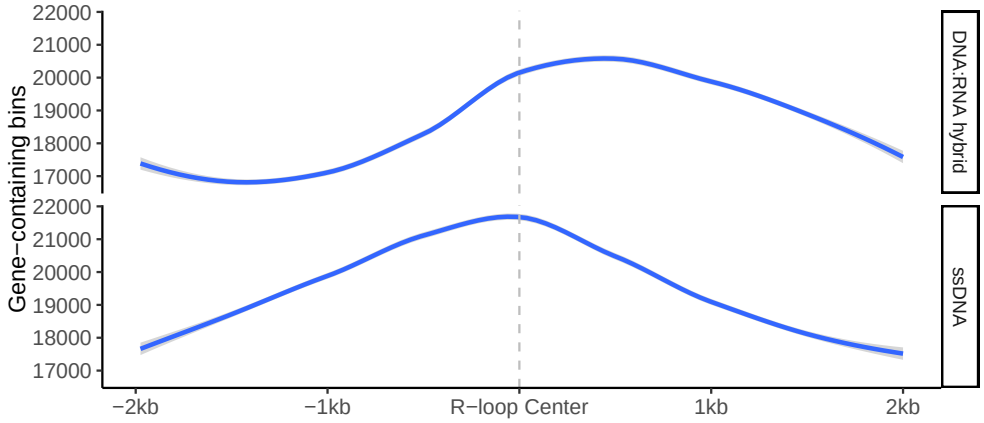

Figure 4.67 Gene distribution on *Arabidopsis* R-loops. The regions around R-loop centers were compared with protein-coding genes and overlaps were counted.

5. DISCUSSION

In this thesis, the potential relationship between R-loops and nucleotide excision repair (NER) was examined. R-loop-forming locations on the genome were mapped by several different methods including DRIP-seq, ssDRIP-seq, qDRIP-seq and RR-ChIP-seq, previously. These methods, in general, rely on two different approaches: capturing the R-loops by S9.6 antibody or by a mutated version of RNase H1 enzyme which can bind but cannot resolve R-loops (Crossley et al., 2019,2; Hamperl et al., 2017; Tan-Wong et al., 2019; Yang et al., 2019). We have tested the R-loops obtained by ssDRIP-seq, qDRIP-seq and RR-ChIP-seq methods from HeLa cells and found discrepancies between them. In the context of genome coverage, chromatin accessibility and histone marker occupancy, the R-loops from different methods differentiated from each other. While R-loops from ssDRIP-seq data were more homogeneous across the genome, other methods showed significant peaks on some regions of the genome. Furthermore, we assessed these three sets of R-loops in terms of UV-induced damage and NER efficiency profiles and the results also revealed no similarity between the datasets. The inconsistencies between the genomic distributions of R-loops of S9.6-based and RNase H-based methods were mentioned in a previous study (Miller et al., 2022). They also stated that the R-loops mapped by S9.6-based methods covered a larger portion of the genome than the R-loops from RNase H-based methods, which altogether could explain why we observed distinct genome coverage patterns.

Due to these inconsistencies, we decided to test R-loop locations from two databases that were more recently published: R-loopBase and RLBase (Lin et al., 2022; Miller et al., 2022). The R-loops from these databases were more consistent with each other in terms of chromatin accessibility and position relative to genes. On the other hand, the damage and repair profiles did not exactly matched, leading us to choose RLBase database for further analysis since the R-loops from RLBase database are the consensus R-loops from many cell lines that passed a set of quality-control criteria (Miller et al., 2022).

The UV-induced CPD damage formation tendency on the nucleic acid structures de-

pends on the distance and the dihedral angle between the C5-C6 double bonds on the aromatic carbon rings of adjacent pyrimidines (Mao et al., 2018; Nayis et al., 2023; Stark et al., 2022). Our observation of CPD accumulation levels being different on ssDNA and DNA:RNA hybrid strands of R-loops and the dsDNA surrounding them led us to hypothesize that this difference could be due to the structural distinctions that result in varying distances and angles between the adjacent pyrimidines that ease or prevent CPD formation. Molecular dynamics simulations of these nucleic acid structures with the same sequence revealed that the ssDNA harbours distance and angle fluctuations between the adjacent thymines when compared to the same thymines in DNA:RNA hybrid and dsDNA, probably because of the high motility of nucleotides in ssDNA due to the lack of restricting hydrogen bonds. On the other hand, the distance and dihedral angles were the lowest between the same adjacent thymines of DNA:RNA hybrid structure, which could be due to the exceptional stability of DNA:RNA hybrid structures (Hegazy et al., 2020; Ratmeyer, Vinayak, Zhong, Zon & Wilson, 1994). The consistency of distance and angle properties with the damage levels supported our hypothesis and suggested that thymine dimers in ssDNA less frequently align appropriately which make the CPD formation less possible while the distances and angles between thymine dimers of DNA:RNA hybrid are more convenient for CPD formation when exposed to UV.

The first glance at the NER profiles on R-loops suggested a more efficient repair on both strands when compared to the surrounding dsDNA. Repair efficiency was higher on DNA:RNA hybrid than on ssDNA. These results were observed both in 12-minute and 1-hour time points after UV exposure with a more pronounced strand difference at 1-hour due to the repair preference of transcription-coupled NER (TC-NER) which should be fully active at 1-hour time-point but not at 12-minute time-point (Hu et al., 2017). In addition, the same observations were present on the *Arabidopsis* R-loops even though we could not eliminate the dependence of NER activity to damage abundance due to the absence of Damage-seq data. Moreover, cytosine-to-thymine (C->T) mutation levels were lower on the R-loops in melanoma and skin adenocarcinoma genomes supporting the higher repair efficiency results while skin cutaneous melanoma genome showed the opposite pattern which could be due to differentiations between cancer subtypes. In addition, the mutation counts on R-loop-containing regions were lower than expected by their sequence content. The reason why we observed a more efficient repair and less mutations on R-loops could be R-loops being on highly expressed regions on the genome which would have a higher number of stalled RNA polymerases and attract TC-NER proteins more efficiently. Consistently, R-loops are mostly found on transcriptionally active sites, where the TC-NER is expected to be high (Canturk et al., 2016; Castillo-Guzman

& Chédin, 2021). However, our analysis showed that when the TC-NER activity was inhibited or only the regions with the same expression levels were tested, repair efficiency on R-loops was still higher than the surrounding regions and the mutations on R-loop regions were still significantly lower. Similarly, in *Arabidopsis*, the higher repair on R-loops remained when TC-NER was inhibited. These findings indicated that the high efficiency of repair on R-loops was not due to high expression leading to high TC-NER activity. Instead, other factors should be present that enhance NER efficiency on R-loops.

R-loops are tightly regulated by various proteins that can act as helicases, nucleases or prevent the factors that ease R-loop formation (Hegazy et al., 2020). To investigate if subsets R-loops that are regulated by different combinations of regulators acquire different repair efficiencies, we divided human genome into states according to the R-loop regulator binding events using Hidden Markov Models (HMMs). The analysis of R-loops from the 10 states did not reveal different repair or damage profiles in correlation with the regulator combinations. Instead, the repair and damage profiles differed between the states with high and low regulator occupancy. R-loops from the states with higher regulator binding showed higher repair efficiency when compared to the R-loops from the states with lower protein binding and this difference did not change in the TC-NER-inactivated cells eliminating the doubt that the cause of higher repair efficiency was the TC-NER activity. Furthermore, the damage abundance was lower on both strands than the surroundings of the R-loops from the states with high protein binding while other R-loops resembled the general profile which showed higher damage on DNA:RNA hybrid and lower damage on ssDNA. The states with higher protein binding also had high chromatin accessibilities; therefore, the correlation between damage formation, repair efficiency, protein binding frequency and chromatin accessibility suggested that the high repair efficiency observed on R-loops was not solely due to the presence R-loop structure. High chromatin accessibility could increase the repair efficiency by easing the access of proteins of NER pathway as proteins have to surpass histones and other chromatin-binding factor on heterochromatin regions (Klemm, Shipony & Greenleaf, 2019). In addition, the slight drop of repair efficiency at the R-loop centers which was observed in the general profile on all R-loops was also observed on the R-loop of the states with high protein occupancy, indicating that this drop might be due to the reduced access of repair proteins to R-loop centers that were occupied by the regulator binding, which was previously observed with the transcription factor (TF) binding (Frigola et al., 2021; Sabarinathan, Mularoni, Deu-Pons, Gonzalez-Perez & López-Bigas, 2016). Similarly, the bound proteins at the time of UV exposure might have protected the R-loops from CPD formation on the highly occupied states as

previously reported at the TF-binding sites (Frigola et al., 2021).

6. CONCLUSION

Research to understand more about R-loops is important because of their involvement in many cellular processes as well as being threats to genome integrity upon insufficient regulation. Our study pointed out the inconsistencies between the R-loop datasets potentially resulting from the accuracy and sensitivity of the methods and the dynamic nature of the R-loops which led us choose the most recent and quality-filtered set of R-loops for further analysis.

CPD damage abundance differed on the ssDNA and DNA:RNA hybrid strands of the R-loops and the dsDNA surrounding them. This difference was explained by the dihedral angle between the C5-C6 double bonds and the distance between the C5s and C6s of the adjacent thymines, using molecular dynamics simulations. The DNA:RNA hybrid structure which displayed the lowest angle and distance throughout the trajectory accumulated the highest damage, since lower angles and distances are more favorable for CPD formation as stated in the literature. Consistently, ssDNA which displayed the highest angles as well as fluctuating distance and angles accumulated the lowest damage. We provided another evidence for the phenomena that lower angles and distances being more favorable for the CPD formation. We also demonstrated that the presence of DNA:RNA hybrids may increase the sensitivity for UV-induced damage while ssDNA structures protect genome from CPD formation.

Efficiency of CPD repair by nucleotide excision repair (NER) was higher on R-loops than their surroundings. This observation was also supported by the lower accumulation of cytosine-to-thymine mutations in melanoma and skin adenocarcinoma genomes. However, this difference was not due to higher transcription-coupled nucleotide excision repair (TC-NER) on R-loops since higher global genome nucleotide excision repair (GG-NER) on R-loops was still observed in TC-NER-inactivated cells. Instead, seeing this profile only on genomic states with high chromatin accessibility revealed that R-loops are more efficiently repaired because they are mostly located on open chromatin where NER proteins can access more easily. Furthermore, high regulator binding slightly altered the access of NER proteins at R-loop

centers while also reducing the CPD damage formation on the same regions.

BIBLIOGRAPHY

(2019). Picard toolkit. `https://broadinstitute.github.io/picard/`.

Aboussekhra, A. & Thoma, F. (1999). Tata-binding protein promotes the selective formation of uv-induced (6-4)-photoproducts and modulates dna repair in the tata box. *The EMBO journal*, *18*(2), 433–443.

Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., & Lindahl, E. (2015). Gromacs: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, *1*, 19–25.

Akkose, U. & Adebali, O. (2023). Boquila: Ngs read simulator to eliminate read nucleotide bias in sequence analysis. *Turkish Journal of Biology*, *47*(2), 141–157.

Al-Hadid, Q. & Yang, Y. (2016). R-loop: an emerging regulator of chromatin dynamics. *Acta biochimica et biophysica Sinica*, *48*(7), 623–631.

Andersson, R., Refsing Andersen, P., Valen, E., Core, L. J., Bornholdt, J., Boyd, M., Heick Jensen, T., & Sandelin, A. (2014). Nuclear stability and transcriptional directionality separate functionally distinct rna species. *Nature communications*, *5*(1), 5336.

Arab, K., Karaulanov, E., Musheev, M., Trnka, P., Schäfer, A., Grummt, I., & Niehrs, C. (2019). Gadd45a binds r-loops and recruits tet1 to cpg island promoters. *Nature genetics*, *51*(2), 217–223.

Becker, J. S., Nicetto, D., & Zaret, K. S. (2016). H3k9me3-dependent heterochromatin: barrier to cell fate changes. *Trends in Genetics*, *32*(1), 29–41.

Benayoun, B. A., Pollina, E. A., Ucar, D., Mahmoudi, S., Karra, K., Wong, E. D., Devarajan, K., Daugherty, A. C., Kundaje, A. B., Mancini, E., et al. (2014). H3k4me3 breadth is linked to cell identity and transcriptional consistency. *Cell*, *158*(3), 673–688.

Berendsen, H. J. (1986). Practical algorithms for dynamic simulations. *Molecular-dynamics simulation of statistical-mechanical systems*.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, *28*(1), 235–242.

Bohm, K. A., Morledge-Hampton, B., Stevison, S., Mao, P., Roberts, S. A., & Wyrick, J. J. (2023). Genome-wide maps of rare and atypical uv photoproducts reveal distinct patterns of damage formation and mutagenesis in yeast chromatin. *Proceedings of the National Academy of Sciences*, *120*(10), e2216907120.

Bou-Nader, C., Bothra, A., Garboczi, D. N., Leppla, S. H., & Zhang, J. (2022). Structural basis of r-loop recognition by the s9. 6 monoclonal antibody. *Nature Communications*, *13*(1), 1641.

Brickner, J. R., Garzon, J. L., & Cimprich, K. A. (2022). Walking a tightrope: The complex balancing act of r-loops in genome stability. *Molecular cell*.

Bryan, D. S., Ransom, M., Adane, B., York, K., & Hesselberth, J. R. (2014). High resolution mapping of modified dna nucleobases using excision repair enzymes. *Genome research*, *24*(9), 1534–1542.

Bussi, G., Donadio, D., & Parrinello, M. (2007). Canonical sampling through veloc-

ity rescaling. *The Journal of chemical physics*, *126*(1).

Cadet, J., Mouret, S., Ravanat, J.-L., & Douki, T. (2012). Photoinduced damage to cellular dna: direct and photosensitized reactions. *Photochemistry and Photobiology*, *88*(5), 1048–1065.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). Blast+: architecture and applications. *BMC bioinformatics*, *10*, 1–9.

Canturk, F., Karaman, M., Selby, C. P., Kemp, M. G., Kulaksiz-Erkmen, G., Hu, J., Li, W., Lindsey-Boltz, L. A., & Sancar, A. (2016). Nucleotide excision repair by dual incisions in plants. *Proceedings of the National Academy of Sciences*, *113*(17), 4706–4710.

Castellano-Pozo, M., Santos-Pereira, J. M., Rondón, A. G., Barroso, S., Andújar, E., Pérez-Alegre, M., García-Muse, T., & Aguilera, A. (2013). R loops are linked to histone h3 s10 phosphorylation and chromatin condensation. *Molecular cell*, *52*(4), 583–590.

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

Chakraborty, P. & Grosse, F. (2011). Human dhx9 helicase preferentially unwinds rna-containing displacement loops (r-loops) and g-quadruplexes. *DNA repair*, *10*(6), 654–665.

Chakraborty, P., Huang, J. T., & Hiom, K. (2018). Dhx9 helicase promotes r-loop formation in cells with impaired rna splicing. *Nature communications*, *9*(1), 4346.

Cheatham III, T. E. & Case, D. A. (2013). Twenty-five years of nucleic acid simulations. *Biopolymers*, *99*(12), 969–977.

Chen, L., Chen, J.-Y., Zhang, X., Gu, Y., Xiao, R., Shao, C., Tang, P., Qian, H., Luo, D., Li, H., et al. (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, S., Zhou, Y., Chen, Y., & Gu, J. (2018). fastp: an ultra-fast all-in-one fastq preprocessor. *Bioinformatics*, *34*(17), i884–i890.

Cheng, L., Wang, W., Yao, Y., & Sun, Q. (2021). Mitochondrial rnase h1 activity regulates r-loop homeostasis to maintain genome integrity and enable early embryogenesis in arabidopsis. *PLoS Biology*, *19*(8), e3001357.

Chimera, U. (2004). a visualization system for exploratory research and analysis. pettersen ef, goddard td, huang cc, couch gs, greenblatt dm, meng ec, ferrin te. *J Comput Chem*, *25*(13), 1605–12.

Cloutier, S. C., Wang, S., Ma, W. K., Al Husini, N., Dhoondia, Z., Ansari, A., Pascuzzi, P. E., & Tran, E. J. (2016). Regulated formation of lncrna-dna hybrids enables faster transcriptional induction and environmental adaptation. *Molecular cell*, *61*(3), 393–404.

Compe, E. & Egly, J.-M. (2012). Tfiih: when transcription met dna repair. *Nature reviews Molecular cell biology*, *13*(6), 343–354.

Consortium, E. P. et al. (2012). An integrated encyclopedia of dna elements in the human genome. *Nature*, *489*(7414), 57.

Costantino, L. & Koshland, D. (2015). The yin and yang of r-loop biology. *Current opinion in cell biology*, *34*, 39–45.

Costantino, L. & Koshland, D. (2018). Genome-wide map of r-loop-induced damage

reveals how a subset of r-loops contributes to genomic instability. *Molecular cell*, *71*(4), 487–497.

Crossley, M. P., Bocek, M., & Cimprich, K. A. (2019). R-loops as cellular regulators and genomic threats. *Molecular cell*, *73*(3), 398–411.

Crossley, M. P., Bocek, M. J., Hamperl, S., Swigut, T., & Cimprich, K. A. (2020). qdrip: a method to quantitatively assess rna–dna hybrid formation genome-wide. *Nucleic acids research*, *48*(14), e84–e84.

Darden, T., York, D., & Pedersen, L. (1993). Particle mesh ewald: An n log (n) method for ewald sums in large systems. *The Journal of chemical physics*, *98*(12), 10089–10092.

De Laat, W. L., Appeldoorn, E., Sugasawa, K., Weterings, E., Jaspers, N. G., & Hoeijmakers, J. H. (1998). Dna-binding polarity of human replication protein a positions nucleases in nucleotide excision repair. *Genes & development*, *12*(16), 2598.

de Laat, W. L., Jaspers, N. G., & Hoeijmakers, J. H. (1999a). Molecular mechanism of nucleotide excision repair. *Genes & development*, *13*(7), 768–785.

de Laat, W. L., Jaspers, N. G., & Hoeijmakers, J. H. (1999b). Molecular mechanism of nucleotide excision repair. *Genes & development*, *13*(7), 768–785.

DeLano, W. L. et al. (2002). Pymol: An open-source molecular graphics tool. *CCP4 Newsl. Protein Crystallogr*, *40*(1), 82–92.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). Star: ultrafast universal rna-seq aligner. *Bioinformatics*, *29*(1), 15–21.

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

Ernst, J. & Kellis, M. (2017). Chromatin-state discovery and genome annotation with chromhmm. *Nature protocols*, *12*(12), 2478–2492.

Frigola, J., Sabarinathan, R., Gonzalez-Perez, A., & Lopez-Bigas, N. (2021). Variable interplay of uv-induced dna damage and repair at transcription factor binding sites. *Nucleic Acids Research*, *49*(2), 891–901.

Frigola, J., Sabarinathan, R., Mularoni, L., Muiños, F., Gonzalez-Perez, A., & López-Bigas, N. (2017). Reduced mutation rate in exons due to differential mismatch repair. *Nature genetics*, *49*(12), 1684–1692.

Gan, W., Guan, Z., Liu, J., Gui, T., Shen, K., Manley, J. L., & Li, X. (2011). R-loop-mediated genomic instability is caused by impairment of replication fork progression. *Genes & development*, *25*(19), 2041–2056.

Gatti, V., De Domenico, S., Melino, G., & Peschiaroli, A. (2023). Senataxin and r-loops homeostasis: multifaced implications in carcinogenesis. *Cell Death Discovery*, *9*(1), 145.

Ginno, P. A., Lim, Y. W., Lott, P. L., Korf, I., & 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.

Gómez-González, B. & Aguilera, A. (2020). Looping the (r) loop in dsb repair via rna methylation. *Molecular cell*, *79*(3), 361–362.

Gong, F., Kwon, Y., & Smerdon, M. J. (2005). Nucleotide excision repair in chromatin and the right of entry. *DNA repair*, *4*(8), 884–896.

Graf, M., Bonetti, D., Lockhart, A., Serhal, K., Kellner, V., Maicher, A., Jolivet, P., Teixeira, M. T., & Luke, B. (2017). Telomere length determines terra and

r-loop regulation through the cell cycle. *Cell*, *170*(1), 72–85.

Grant, B. J., Rodrigues, A. P., ElSawy, K. M., McCammon, J. A., & Caves, L. S. (2006). Bio3d: an r package for the comparative analysis of protein structures. *Bioinformatics*, *22*(21), 2695–2696.

Grunseich, C., Wang, I. X., Watts, J. A., Burdick, J. T., Guber, R. D., Zhu, Z., Bruzel, A., Lanman, T., Chen, K., Schindler, A. B., et al. (2018). Senataxin mutation reveals how r-loops promote transcription by blocking dna methylation at gene promoters. *Molecular cell*, *69*(3), 426–437.

Hamperl, S., Bocek, M. J., Saldivar, J. C., Swigut, T., & Cimprich, K. A. (2017). Transcription-replication conflict orientation modulates r-loop levels and activates distinct dna damage responses. *Cell*, *170*(4), 774–786.

Hartono, S. R., Malapert, A., Legros, P., Bernard, P., Chédin, F., & 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.

Hatchi, E., Skourti-Stathaki, K., Ventz, S., Pinello, L., Yen, A., Kamieniarz-Gdula, K., Dimitrov, S., Pathania, S., McKinney, K. M., Eaton, M. L., et al. (2015). Brca1 recruitment to transcriptional pause sites is required for r-loop-driven dna damage repair. *Molecular cell*, *57*(4), 636–647.

Hayward, N. K., Wilmott, J. S., Waddell, N., Johansson, P. A., Field, M. A., Nones, K., Patch, A.-M., Kakavand, H., Alexandrov, L. B., Burke, H., et al. (2017). Whole-genome landscapes of major melanoma subtypes. *Nature*, *545*(7653), 175–180.

Hegazy, Y. A., Fernando, C. M., & Tran, E. J. (2020). The balancing act of r-loop biology: The good, the bad, and the ugly. *Journal of Biological Chemistry*, *295*(4), 905–913.

Hodis, E., Watson, I. R., Kryukov, G. V., Arold, S. T., Imielinski, M., Theurillat, J.-P., Nickerson, E., Auclair, D., Li, L., Place, C., et al. (2012). A landscape of driver mutations in melanoma. *Cell*, *150*(2), 251–263.

Hu, J., Adar, S., Selby, C. P., Lieb, J. D., & Sancar, A. (2015). Genome-wide analysis of human global and transcription-coupled excision repair of uv damage at single-nucleotide resolution. *Genes & development*, *29*(9), 948–960.

Hu, J., Adebali, O., Adar, S., & Sancar, A. (2017). Dynamic maps of uv damage formation and repair for the human genome. *Proceedings of the National Academy of Sciences*, *114*(26), 6758–6763.

Hu, J., Li, W., Adebali, O., Yang, Y., Oztas, O., Selby, C. P., & Sancar, A. (2019). Genome-wide mapping of nucleotide excision repair with xr-seq. *Nature protocols*, *14*(1), 248–282.

Hu, J., Lieb, J. D., Sancar, A., & Adar, S. (2016). Cisplatin dna damage and repair maps of the human genome at single-nucleotide resolution. *Proceedings of the National Academy of Sciences*, *113*(41), 11507–11512.

Huang, Y., Azgari, C., Yin, M., Chiou, Y.-Y., Lindsey-Boltz, L. A., Sancar, A., Hu, J., & Adebali, O. (2022). Effects of replication domains on genome-wide uv-induced dna damage and repair. *PLoS Genetics*, *18*(9), e1010426.

Humphrey, W., Dalke, A., & Schulten, K. (1996). Vmd: visual molecular dynamics. *Journal of molecular graphics*, *14*(1), 33–38.

Hung, K.-F., Sidorova, J. M., Nghiem, P., & Kawasumi, M. (2020). The 6-4 photoproduct is the trigger of uv-induced replication blockage and atr activation.

Proceedings of the National Academy of Sciences, *117*(23), 12806–12816.

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics*, *79*(2), 926–935.

Katoh, K. & Standley, D. M. (2013). Mafft multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution*, *30*(4), 772–780.

Kaya, S., Adebali, O., Oztas, O., & Sancar, A. (2022). Genome-wide excision repair map of cyclobutane pyrimidine dimers in arabidopsis and the roles of csa1 and csa2 proteins in transcription-coupled repair. *Photochemistry and photobiology*, *98*(3), 707–712.

Kemp, M. G., Reardon, J. T., Lindsey-Boltz, L. A., & Sancar, A. (2012). Mechanism of release and fate of excised oligonucleotides during nucleotide excision repair. *Journal of Biological Chemistry*, *287*(27), 22889–22899.

Klemm, S. L., Shipony, Z., & Greenleaf, W. J. (2019). Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics*, *20*(4), 207–220.

Kuhn, R. M., Haussler, D., & Kent, W. J. (2013). The ucsc genome browser and associated tools. *Briefings in bioinformatics*, *14*(2), 144–161.

Langmead, B. & Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature methods*, *9*(4), 357–359.

Law, Y. K., Azadi, J., Crespo-Hernández, C. E., Olmon, E., & Kohler, B. (2008). Predicting thymine dimerization yields from molecular dynamics simulations. *Biophysical journal*, *94*(9), 3590–3600.

Law, Y. K., Forties, R. A., Liu, X., Poirier, M. G., & Kohler, B. (2013). Sequence-dependent thymine dimer formation and photoreversal rates in double-stranded dna. *Photochemical & Photobiological Sciences*, *12*(8), 1431–1439.

Lee, W. & Matsika, S. (2022). Mechanistic aspects of the effect of flanking nucleotide sequence on cpd formation and cpd self-repair in dna. *The Journal of Physical Chemistry B*, *127*(1), 18–25.

Leffell, D. J. (2000). The scientific basis of skin cancer. *Journal of the American Academy of Dermatology*, *42*(1), S18–S22.

Li, C., Wang, H., Yin, Z., Fang, P., Xiao, R., Xiang, Y., Wang, W., Li, Q., Huang, B., Huang, J., et al. (2021). Ligand-induced native g-quadruplex stabilization impairs transcription initiation. *Genome research*, *31*(9), 1546–1560.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., & Subgroup, . G. P. D. P. (2009). The sequence alignment/map format and samtools. *bioinformatics*, *25*(16), 2078–2079.

Li, Q., Meissner, T. B., Wang, F., Du, Z., Ma, S., Kshirsagar, S., Tilburgs, T., Buenrostro, J. D., Uesugi, M., & Strominger, J. L. (2021). Elf3 activated by a superenhancer and an autoregulatory feedback loop is required for high-level hla-c expression on extravillous trophoblasts. *Proceedings of the National Academy of Sciences*, *118*(9), e2025512118.

Li, W. & Sancar, A. (2020). Methodologies for detecting environmentally induced dna damage and repair. *Environmental and molecular mutagenesis*, *61*(7), 664–679.

Lin, R., Zhong, X., Zhou, Y., Geng, H., Hu, Q., Huang, Z., Hu, J., Fu, X.-D., Chen, L., & Chen, J.-Y. (2022). R-loopbase: a knowledgebase for genome-wide r-

loop formation and regulation. *Nucleic acids research*, *50*(D1), D303–D315.
Liu, J., Ali, M., & Zhou, Q. (2020). Establishment and evolution of heterochromatin. *Annals of the New York Academy of Sciences*, *1476*(1), 59–77.
Lockhart, A., Pires, V. B., Bento, F., Kellner, V., Luke-Glaser, S., Yakoub, G., Ulrich, H. D., & Luke, B. (2019). Rnase h1 and h2 are differentially regulated to process rna-dna hybrids. *Cell reports*, *29*(9), 2890–2900.
Lu, C., Gutierrez-Bayona, N. E., & Taylor, J.-S. (2021). The effect of flanking bases on direct and triplet sensitized cyclobutane pyrimidine dimer formation in dna depends on the dipyrimidine, wavelength and the photosensitizer. *Nucleic acids research*, *49*(8), 4266–4280.
Lu, W.-T., Hawley, B. R., Skalka, G. L., Baldock, R. A., Smith, E. M., Bader, A. S., Malewicz, M., Watts, F. Z., Wilczynska, A., & Bushell, M. (2018). Drosha drives the formation of dna: Rna hybrids around dna break sites to facilitate dna repair. *Nature communications*, *9*(1), 532.
Mackay, R. P., Xu, Q., & Weinberger, P. M. (2020). R-loop physiology and pathology: a brief review. *DNA and Cell Biology*, *39*(11), 1914–1925.
Mao, P., Brown, A. J., Esaki, S., Lockwood, S., Poon, G. M., Smerdon, M. J., Roberts, S. A., & Wyrick, J. J. (2018). Ets transcription factors induce a unique uv damage signature that drives recurrent mutagenesis in melanoma. *Nature communications*, *9*(1), 2626.
Mao, P., Smerdon, M. J., Roberts, S. A., & Wyrick, J. J. (2016). Chromosomal landscape of uv damage formation and repair at single-nucleotide resolution. *Proceedings of the National Academy of Sciences*, *113*(32), 9057–9062.
Mao, P. & Wyrick, J. J. (2020). Genome-wide mapping of uv-induced dna damage with cpd-seq. *The Nucleus*, 79–94.
Mao, P., Wyrick, J. J., Roberts, S. A., & Smerdon, M. J. (2017). Uv-induced dna damage and mutagenesis in chromatin. *Photochemistry and photobiology*, *93*(1), 216–228.
Maréchal, A. & Zou, L. (2015). Rpa-coated single-stranded dna as a platform for post-translational modifications in the dna damage response. *Cell research*, *25*(1), 9–23.
Marteijn, J. A., Lans, H., Vermeulen, W., & Hoeijmakers, J. H. (2014). Understanding nucleotide excision repair and its roles in cancer and ageing. *Nature reviews Molecular cell biology*, *15*(7), 465–481.
Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, *17*(1), 10–12.
Mathieu, N., Kaczmarek, N., Rüthemann, P., Luch, A., & Naegeli, H. (2013). Dna quality control by a lesion sensor pocket of the xeroderma pigmentosum group d helicase subunit of tfiih. *Current Biology*, *23*(3), 204–212.
Menoni, H., Wienholz, F., Theil, A. F., Janssens, R. C., Lans, H., Campalans, A., Radicella, J. P., Marteijn, J. A., & Vermeulen, W. (2018). The transcription-coupled dna repair-initiating protein csb promotes xrcc1 recruitment to oxidative dna damage. *Nucleic acids research*, *46*(15), 7747–7756.
Miller, H. E., Montemayor, D., Abdul, J., Vines, A., Levy, S. A., Hartono, S. R., Sharma, K., Frost, B., Chédin, F., & Bishop, A. J. (2022). Quality-controlled r-loop meta-analysis reveals the characteristics of r-loop consensus regions. *Nucleic acids research*, *50*(13), 7260–7286.
Monsen, R. C., Trent, J. O., & Chaires, J. B. (2022). G-quadruplex dna: a longer

story. *Accounts of Chemical Research*, *55*(22), 3242–3252.
Morales, J. C., Richard, P., Patidar, P. L., Motea, E. A., Dang, T. T., Manley, J. L., & Boothman, D. A. (2016). Xrn2 links transcription termination to dna damage and replication stress. *PLoS genetics*, *12*(7), e1006107.
Nayis, A., Liebl, K., & Zacharias, M. (2023). Coupling of conformation and cpd damage in nucleosomal dna. *Biophysical Chemistry*, 107050.
Nguyen, H. D., Yadav, T., Giri, S., Saez, B., Graubert, T. A., & 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.
Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A., & Minh, B. Q. (2015). Iq-tree: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular biology and evolution*, *32*(1), 268–274.
Niehrs, C. & Luke, B. (2020). Regulatory r-loops as facilitators of gene expression and genome stability. *Nature reviews Molecular cell biology*, *21*(3), 167–178.
Ohle, C., Tesorero, R., Schermann, G., Dobrev, N., Sinning, I., & Fischer, T. (2016). Transient rna-dna hybrids are required for efficient double-strand break repair. *Cell*, *167*(4), 1001–1013.
Oztas, O., Selby, C. P., Sancar, A., & Adebali, O. (2018). Genome-wide excision repair in arabidopsis is coupled to transcription and reflects circadian gene expression patterns. *Nature communications*, *9*(1), 1503.
Pan, Z., Hariharan, M., Arkin, J. D., Jalilov, A. S., McCullagh, M., Schatz, G. C., & Lewis, F. D. (2011). Electron donor–acceptor interactions with flanking purines influence the efficiency of thymine photodimerization. *Journal of the American Chemical Society*, *133*(51), 20793–20798.
Parrinello, M. & Rahman, A. (1981). Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied physics*, *52*(12), 7182–7190.
Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature methods*, *14*(4), 417–419.
Petermann, E., Lan, L., & Zou, L. (2022). Sources, resolution and physiological relevance of r-loops and rna–dna hybrids. *Nature reviews Molecular cell biology*, *23*(8), 521–540.
Pfeifer, G. P. (2020). Mechanisms of uv-induced mutations and skin cancer. *Genome instability & disease*, *1*(3), 99–113.
Polo, S. E. & Almouzni, G. (2015). Chromatin dynamics after dna damage: the legacy of the access–repair–restore model. *DNA repair*, *36*, 114–121.
Quinlan, A. R. & Hall, I. M. (2010). Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842.
Ramachandran, S., Ma, T. S., Griffin, J., Ng, N., Foskolou, I. P., Hwang, M.-S., Victori, P., Cheng, W.-C., Buffa, F. M., Leszczynska, K. B., et al. (2021). Hypoxia-induced setx links replication stress with the unfolded protein response. *Nature Communications*, *12*(1), 3686.
Ramírez, F., Dündar, F., Diehl, S., Grüning, B. A., & Manke, T. (2014). deeptools: a flexible platform for exploring deep-sequencing data. *Nucleic acids research*, *42*(W1), W187–W191.
Ratmeyer, L., Vinayak, R., Zhong, Y. Y., Zon, G., & Wilson, W. D. (1994). Sequence specific thermodynamic and structural properties for dna.rna duplexes.

Biochemistry, *33*(17), 5298–5304.

Reardon, J. T. & Sancar, A. (2005). Nucleotide excision repair. *Progress in nucleic acid research and molecular biology*, *79*, 183–235.

Rinaldi, C., Pizzul, P., Longhese, M. P., & Bonetti, D. (2021). Sensing r-loop-associated dna damage to safeguard genome stability. *Frontiers in Cell and Developmental Biology*, *8*, 618157.

Rowley, M. J. & Corces, V. G. (2018). Organizational principles of 3d genome architecture. *Nature Reviews Genetics*, *19*(12), 789–800.

Rühle, V. (2008). Pressure coupling/barostats. *Journal Club*, 1–5.

Sabarinathan, R., Mularoni, L., Deu-Pons, J., Gonzalez-Perez, A., & López-Bigas, N. (2016). Nucleotide excision repair is impaired by binding of transcription factors to dna. *Nature*, *532*(7598), 264–267.

Sahakyan, A. B., Chambers, V. S., Marsico, G., Santner, T., Di Antonio, M., & Balasubramanian, S. (2017). Machine learning model for sequence-driven dna g-quadruplex formation. *Scientific reports*, *7*(1), 14535.

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

Schärer, O. D. (2013). Nucleotide excision repair in eukaryotes. *Cold Spring Harbor perspectives in biology*, *5*(10), a012609.

Scrima, A., Koníčková, R., Czyzewski, B. K., Kawasaki, Y., Jeffrey, P. D., Groisman, R., Nakatani, Y., Iwai, S., Pavletich, N. P., & Thomä, N. H. (2008). Structural basis of uv dna-damage recognition by the ddb1–ddb2 complex. *Cell*, *135*(7), 1213–1223.

Sequeira-Mendes, J., Aragüez, I., Peiró, R., Mendez-Giraldez, R., Zhang, X., Jacobsen, S. E., Bastolla, U., & Gutierrez, C. (2014). The functional topography of the arabidopsis genome is organized in a reduced number of linear motifs of chromatin states. *The Plant Cell*, *26*(6), 2351–2366.

Shivji, M. K., Renaudin, X., Williams, Ç. H., & Venkitaraman, A. R. (2018). Brca2 regulates transcription elongation by rna polymerase ii to prevent r-loop accumulation. *Cell reports*, *22*(4), 1031–1039.

Showalter, S. A. & Brüschweiler, R. (2007). Validation of molecular dynamics simulations of biomolecules using nmr spin relaxation as benchmarks: application to the amber99sb force field. *Journal of chemical theory and computation*, *3*(3), 961–975.

So, C. C. & Martin, A. (2019). Dsb structure impacts dna recombination leading to class switching and chromosomal translocations in human b cells. *PLoS Genetics*, *15*(4), e1008101.

Sollier, J. & Cimprich, K. A. (2015). Breaking bad: R-loops and genome integrity. *Trends in cell biology*, *25*(9), 514–522.

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

Song, C., Hotz-Wagenblatt, A., Voit, R., & Grummt, I. (2017). Sirt7 and the dead-box helicase ddx21 cooperate to resolve genomic r loops and safeguard genome stability. *Genes & development*, *31*(13), 1370–1381.

Stark, B., Poon, G. M., & Wyrick, J. J. (2022). Molecular mechanism of uv dam-

age modulation in nucleosomes. *Computational and Structural Biotechnology Journal*, *20*, 5393–5400.

Stavnezer, J. & Schrader, C. E. (2006). Mismatch repair converts aid-instigated nicks to double-strand breaks for antibody class-switch recombination. *TRENDS in Genetics*, *22*(1), 23–28.

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

Sugasawa, K. (2008). Xeroderma pigmentosum genes: functions inside and outside dna repair. *Carcinogenesis*, *29*(3), 455–465.

Sun, Q., Csorba, T., Skourti-Stathaki, K., Proudfoot, N. J., & Dean, C. (2013). R-loop stabilization represses antisense transcription at the arabidopsis flc locus. *Science*, *340*(6132), 619–621.

Sun, Z., Zhang, Y., Jia, J., Fang, Y., Tang, Y., Wu, H., & Fang, D. (2020). H3k36me3, message from chromatin to dna damage repair. *Cell & bioscience*, *10*(1), 1–9.

Systèmes, D. (2020). Biovia workbook.

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

Topolska-Woś, A. M., Sugitani, N., Cordoba, J. J., Le Meur, K. V., Le Meur, R. A., Kim, H. S., Yeo, J.-E., Rosenberg, D., Hammel, M., Schärer, O. D., et al. (2020). A key interaction with rpa orients xpa in ner complexes. *Nucleic acids research*, *48*(4), 2173–2188.

Tornaletti, S. & Pfeifer, G. P. (1995). Uv light as a footprinting agent: modulation of uv-induced dna damage by transcription factors bound at the promoters of three human genes. *Journal of molecular biology*, *249*(4), 714–728.

Tu, J., Duan, M., Liu, W., Lu, N., Zhou, Y., Sun, X., & Lu, Z. (2021). Direct genome-wide identification of g-quadruplex structures by whole-genome resequencing. *Nature communications*, *12*(1), 6014.

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., & Berendsen, H. J. (2005a). Gromacs: fast, flexible, and free. *Journal of computational chemistry*, *26*(16), 1701–1718.

Van Der Spoel, D., Lindahl, E., Hess, B., Groenhof, G., Mark, A. E., & Berendsen, H. J. (2005b). Gromacs: fast, flexible, and free. *Journal of computational chemistry*, *26*(16), 1701–1718.

Villarreal, O. D., Mersaoui, S. Y., Yu, Z., Masson, J.-Y., & Richard, S. (2020). Genome-wide r-loop analysis defines unique roles for ddx5, xrn2, and prmt5 in dna/rna hybrid resolution. *Life Science Alliance*, *3*(10).

Waskom, M. L. (2021). Seaborn: statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021.

Wickham, H. (2011). ggplot2. *Wiley interdisciplinary reviews: computational statistics*, *3*(2), 180–185.

Wiedemann, E.-M., Peycheva, M., & Pavri, R. (2016). Dna replication origins in immunoglobulin switch regions regulate class switch recombination in an r-loop-dependent manner. *Cell reports*, *17*(11), 2927–2942.

Xu, C., Wu, Z., Duan, H.-C., Fang, X., Jia, G., & Dean, C. (2021). R-loop resolution promotes co-transcriptional chromatin silencing. *Nature Communications*, *12*(1), 1790.

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

Yang, X., Liu, Q.-L., Xu, W., Zhang, Y.-C., Yang, Y., Ju, L.-F., Chen, J., Chen, Y.-S., Li, K., Ren, J., et al. (2019). m6a promotes r-loop formation to facilitate transcription termination. *Cell research*, *29*(12), 1035–1038.

Yu, K., Chedin, F., Hsieh, C.-L., Wilson, T. E., & 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, C., Guo, H., Zhang, J., Guo, G., Schumaker, K. S., & Guo, Y. (2010). Arabidopsis cockayne syndrome a-like proteins 1a and 1b form a complex with cullin4 and damage dna binding protein 1a and regulate the response to uv irradiation. *The Plant Cell*, *22*(7), 2353–2369.

Zhang, X., Chiang, H.-C., Wang, Y., Zhang, C., Smith, S., Zhao, X., Nair, S. J., Michalek, J., Jatoi, I., Lautner, M., et al. (2017). Attenuation of rna polymerase ii pausing mitigates brca1-associated r-loop accumulation and tumorigenesis. *Nature communications*, *8*(1), 15908.

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W., et al. (2008). Model-based analysis of chip-seq (macs). *Genome biology*, *9*(9), 1–9.

Zhang, Z. Z., Pannunzio, N. R., Han, L., Hsieh, C.-L., Yu, K., & Lieber, M. R. (2014). The strength of an ig switch region is determined by its ability to drive r loop formation and its number of wgcw sites. *Cell reports*, *8*(2), 557–569.

Zhao, H., Zhu, M., Limbo, O., & Russell, P. (2018). Rnase h eliminates r-loops that disrupt dna replication but is nonessential for efficient dsb repair. *EMBO reports*, *19*(5), e45335.

Zheng, K.-w., Zhang, J.-y., He, Y.-d., Gong, J.-y., Wen, C.-j., Chen, J.-n., Hao, Y.-h., Zhao, Y., & Tan, Z. (2020). Detection of genomic g-quadruplexes in living cells using a small artificial protein. *Nucleic acids research*, *48*(20), 11706–11720.

Zheng, Y., Lorenzo, C., & Beal, P. A. (2017). Dna editing in dna/rna hybrids by adenosine deaminases that act on rna. *Nucleic acids research*, *45*(6), 3369–3377.

Zhou, J., Lei, X., Shafiq, S., Zhang, W., Li, Q., Li, K., Zhu, J., Dong, Z., He, X.-j., & Sun, Q. (2023). Ddm1-mediated r-loop resolution and h2a. z exclusion facilitates heterochromatin formation in arabidopsis. *Science Advances*, *9*(32), eadg2699.

Zhu, J., Liu, M., Liu, X., & Dong, Z. (2018). Rna polymerase ii activity revealed by gro-seq and pnet-seq in arabidopsis. *Nature plants*, *4*(12), 1112–1123.