UNIVERSITY OF CALIFORNIA

RIVERSIDE

Biological Studies of Alkylated 2′-Deoxyguanosine Lesions and R-loops

A Dissertation submitted in partial satisfaction

of the requirements for the degree of

Doctor of Philosophy

in

Chemistry

By

Yinan Wang

December 2023

Dissertation Committee:

Dr. Yinsheng Wang, Chairperson

Dr. Wenwan Zhong

Dr. Quan Cheng

The Dissertation of Yinan Wang is approved:

Committee Chairperson

University of California, Riverside

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor, Professor Yinsheng Wang, for having faith in me. I would never come this far without his guidance and support. He always welcomes discussions and provides timely help. From fundamental knowledge in mass spectrometry to cutting-edge research, from academic writing to work ethics, over the past few years, I have learned from him not only knowledge but also how to be a good scientist. His passion and dedication have always inspired me, both in research and personal growth.

I would also like to thank Dr. Quan Cheng and Dr. Wenwan Zhong for serving on my dissertation committee. I sincerely appreciate them for taking the time to provide feedback and suggestions on my progress and development during my time at the University of California, Riverside (UCR). I am grateful for the support from my labmates and staff at UCR. I am proud that I was learning and working in such a fantastic research group. Everyone in the group gave me generous assistance during these years. Especially, I would like to thank Dr. Feng Tang, who guided me in my research and career development for the past five years. I also want to thank Dr. Xiaomei He and Dr. Jun Yuan for their collaborative work, and their efforts contributed greatly to the completion of my thesis. I wish to extend my acknowledgment to Dr. Jiabin Wu and Dr. Hua Du, who taught me to conduct *in vivo* replication and transcription assay.

I am also grateful that my research was supported by funding from the National Institutes of Health and fellowships from the Chemistry graduate program of the University of California Riverside. Finally, I want to express my gratitude to my family members, who are the first and most important motivation for my studying at UCR. Their unconditional love and support gave me the power and confidence to overcome any barrier. I love you so much.

COPYRIGHT ACKNOWLEDGEMENTS

The text and figures in Chapter 2, in part or in full, are a reprint of the material as it appears in *Chem. Res. Toxicol*. 2021, 34, 3, 695–698. The co-author (Dr. Yinsheng Wang) listed in that publication directed and supervised the research which forms the basis of this Chapter. The co-authors (Dr. Jun Wu, and Dr. Jiabin Wu) listed in that publication contributed to the synthesis of N^2-alkyl-dG oligomers.

ABSTRACT OF THE DISSERTATION

Biological Studies of Alkylated 2′-Deoxyguanosine Lesions and R-loops

By

Yinan Wang

Doctor of Philosophy, Graduate Program in Chemistry
University of California, Riverside, December 2023
Dr. Yinsheng Wang, Chairperson

Endogenous and environmental agents can cause DNA damage in cells. A number of alkylating agents introduce modifications into DNA bases that can be cytotoxic and mutagenic to the cell. N^2-alkyl-2'-deoxyguanosine (dG) lesions are a primary type of DNA damage products resulting from alkylation that have been identified in the genome. While most of the DNA damage can be removed through different types of DNA repair pathways, any unrepaired DNA lesions may obstruct replication and transcription processes and further trigger genome instability. However, there is limited understanding regarding the impact of N^2-alkyl-dG lesions in cellular processes and the functions of translesion synthesis (TLS) polymerases in modulating genome instability.

In Chapter 2, we synthesized ODNs harboring N^2-alkyldG lesions and incorporated them into plasmid vectors. By performing *in vivo* replication assay in wild-type and TLS polymerase-deficient *E. coli* cells, we found that N^2-Et-dG is slightly less blocking to DNA replication than the two isomeric N^2-*n*Bu-dG lesions, which display very similar replication bypass efficiencies.

Additionally, Pol II and, to a lesser degree, Pol IV and Pol V are required for the efficient bypass of the N^2-alkyl-dG adducts, and none of these lesions were mutagenic.

In Chapter 3, we used two independent approaches (proximity labeling and affinity pull-down) followed by LC-MS/MS analysis to profile Pol κ-interacting proteins at the proteome-wide level. Our result revealed that Pol κ interacts with DDX23, and Pol κ ChIP-Seq analyses indicated that Pol κ is enriched at R-loop structure loci in chromatin. In addition, Pol κ recruits DDX23 to R-loop sites in chromatin and promotes unwinding of R-loop structure. We also observed an augmented accumulation of R-loops in Pol κ- and DDX23-deficient cells. Together, we discovered an interaction between Pol κ and DDX23, and revealed that the interaction facilitates resolution of R-loops in cells.

In Chapter 4, we demonstrated N^2-alkyl-dG adducts induce R-loop structure accumulation on human chromatin by conducting immunofluorescence microscopy and sp-KAS-seq analysis. R-ChIP-qPCR assay further illustrates R-loop structures formed on lesion insertion sites in lesion-bearing episomal plasmids and cellular transcription assay results showed the lesion-induced R-loops strongly impede transcription. We also observed the accumulation of double-strand breaks (DSBs) after N^2-alkyl-dG incorporation in DDX23-deficient cells. Together, we conclude that N^2-alkyl-dG adducts trigger R-loop accumulation and further induce genome instability in human chromatin.

Together, the research described in this dissertation reveals the adverse biological consequences of N^2-alkyl-dG lesions and provides a much better understanding of the functions of TLS polymerases in modulating replicative bypass of N^2-alkyl-dG lesions and maintaining genome stability.

Table of Contents

Chapter 1. General Introduction 1

DNA Damage and its biological consequences 1

R-loop structures 4

Translesion Synthesis and DNA polymerases 7

Methods 10

In vivo replication and transcription experiments 10

Interaction Proteomics 14

Next-generation Sequencing (NGS) 20

Scope of this Dissertation 25

References 27

Chapter 2. DNA Polymerase II Supports the Replicative Bypass of N2-Alkyl-2′-deoxyguanosine Lesions in Escherichia coli Cells 34

Introduction 34

Methods and Materials 36

Results 39

Discussion 40

Figures 43

References 51

Chapter 3. Polymerase κ Recruits DDX23 to Promote R-loop Resolution 58

Introduction 58

Materials and Methods 59

Results 73

Discussion 80

Figures and Table 77

References 109

Chapter 4. N^2-alkyl-dG lesions trigger R-loop formation in genome.....115

Introduction..... 115

Methods and Materials..... 116

Results..... 123

Discussion..... 129

Figures and Table..... 132

References..... 145

Chapter 5. Concluding Remarks and Future Directions 149

List of Figures

Figure 1-1. Structures of 2′-deoxyguanosine (dG) and N^2-alkylguanine lesions..... 3

Figure 1-2. R-loop removal is achieved by RNase H enzymes that degrade the RNA moiety of the hybrid and by DNA-RNA helicases that unwind the hybrid. 7

Figure 1-3. Two models for lesion bypass by TLS Pols. 9

Figure 1-4. A schematic diagram illustrating the experimental procedures of CRAB assay. 12

Figure 1-5. A schematic diagram illustrating the procedures for the construction of plasmids harboring a site-specifically incorporated lesion. 13

Figure 1-6. A schematic diagram illustrating the CTAB assay system (93). 14

Figure 1-7. Experimental workflows for bottom-up proteomic approaches. 17

Figure 1-8. General workflow of proximity labeling followed by mass spectrometry with peroxidase. 18

Figure 1-9. General workflow of CRISPR-KI followed by mass spectrometry with peroxidase. 20

Figure 1-10. A schematic workflow of Sanger sequencing. 21

Figure 1-11. A schematic diagram of NGS. 22

Figure 1-12. An outline of ChIP-seq workflow. 23

Figure 1-13. A schematic illustration of the principle that spKAS-seq distinguishes R-loops from transcription bubbles. 25

Figure 2-1. ESI-MS & MS/MS characterizations of d(ATGGCGXGCTAT), ... 43

Figure 2-2. ESI-MS & MS/MS characterizations of d(ATGGCGXGCTAT), ... 44

Figure 2-3. Native PAGE (30%) for monitoring the bypass efficiencies and mutation frequencies of N2-alkyl-dG lesions 45

Figure 2-4. Higher-resolution “ultra-zoom scan” ESI-MS of the restriction fragments for the PCR products 46

Figure 2-5. MS/MS for the identification of restriction fragments of PCR products. 47

Figure 2-6. Partial structure of a ternary complex of E. coli DNA polymerase II 48

Figure 2-7. N^2-Alkyl-dG lesions investigated in this study..... 49

Figure 2-8. Restriction digestion and post labeling method for determining the cytotoxic and mutagenic properties of N^2-alkyl-dG lesions 50

Figure 3-1. Western blot analyses of Pol κ-APEX and NLS-APEX expression (a) and labeling (b) in HEK293T cells. 77

Figure 3-2. The distribution of biotinylation sites at different amino acids residues..... 78

Figure 3-3. Biotinylated amino acids residues on lamin B1 and lamin B2 proteins detected with Pol κ-APEX and NLS-APEX. 79

Figure 3-4. Biotinylated amino acids residues on selected proteins detected with Pol κ-APEX. 80

Figure 3-5. Detecting targeted insertion of 3× Flag tag sequence to the C-terminus of Pol κ in HEK293T cells.. 81

Figure 3-6. MS/MS for the $[M+2H]^{2+}$ ion of unique peptides of DDX23 82

Figure 3-7. Detecting targeted insertion of 3× Flag tag sequence to the C-terminus of DDX23 in HEK293T and Pol κ-deficient HEK293T cells. 83

Figure 3-8. Western blot results showing the truncated DDX23 protein input blotted with anti-HA antibody. 84

Figure 3-9. MS/MS for the $[M+3H]^{3+}$ of the biotinylated peptide (IFREDYSITTK) of DDX23 identified from the Pol κ APEX-labeling experiments, where the biotinylated tyrosine residue is highlighted in red. 85

Figure 3-10. Alignment rates and correlation analyses of ChIP-Seq data. 86

Figure 3-11. Pol κ ChIP-Seq signal intensity meta profiles. 87

Figure 3-12. Fluorescence microscopy for analyzing the co-localization of R-loops with human Pol κ and DDX23.. 88

Figure 3-13. Pol κ and R-loop ChIP-Seq results for SNRPN gene in HEK293T cells.... 89

Figure 3-14. Pol κ-ChIP PCR analysis. 90

Figure 3-15. Cell cycle profiles at indicated time points, after release from double thymidine block. 91

Figure 3-16. Alignment rates and correlation analyses of R-ChIP-Seq data. 92

Figure 3-17. The genomic distribution of R-loop loci obtained from R-ChIP-Seq in *POLK*$^{-/-}$ and *DDX23*$^{-/-}$ HEK293T cells. 93

Figure 3-18. APEX labeling and affinity pull-down for revealing the Pol κ interactome. 94

Figure 3-19. Immunoprecipitation followed by Western blot for assessing the interaction between DDX23 and Pol κ. 95

Figure 3-20. ChIP-seq revealed enrichment of Pol κ at R-loop structure loci in chromatin. 96

Figure 3-21. DDX23 is co-localized with Pol κ and R-loop structures. 97

Figure 3-22. R-ChIP-seq revealed increased accumulation of R-loops in DDX23- and Pol κ-deficient HEK293T cells. 98

Figure 3-23. R-loop structures are accumulated in *POLK*$^{-/-}$ and *DDX23*$^{-/-}$ cells. 99

Figure 3-24. Fluorescence imaging analysis revealed increased accumulation of R-loops in the DDX23- and Pol κ-deficient HEK293T cells. 100

Figure 3-25. A proposed model for Pol κ's function in promoting R-loop resolution through DDX23. 101

Figure 4-1. Scheme of N^2-heptynyl-dG synthesis. 127

Figure 4-2. ESI-MS (A) and MS/MS (B) analysis of N^2-heptynyl-dG. 128

Figure 4-3. N^2-alkyl-dG lesions induce elevated R-loop formation. 129

Figure 4-4. spKAS-seq revealed increased accumulation of R-loops in the N^2-heptynyl-dG treated HEK293T cells.. 130

Figure 4-5. R-loops are accumulated at N^2-alkyl-dG incorporation site. 131

Figure 4-6. Scheme of CTAB assay. 132

Figure 4-7. Restriction digestion and postlabeling method for determining the transcriptional bypass efficiencies of N^2-alkyl-dG lesions in HEK293T, *DDX23*$^{-/-}$, and the latter complemented with DDX23. 133

Figure 4-8. Higher resolution 'ultra zoom-scan' ESI-MS for monitoring the $[M-3H]^{3-}$ ions of the restriction fragments of interest. 134

Figure 4-9. LC-MS/MS for monitoring the restriction digestion products of interest corresponding to wild-type or mutated transcripts arising from N^2-alkyl-dG containing substrates in HEK293T cells. 135

Figure 4-10. N^2-alkyl-dG incorporation induces genome instability. 136

Figure 4-11. N^2-heptynyl-dG incorporation assay. 137

Figure 4-12. Representative selected-ion chromatograms (SICs) and MS/MS for monitoring the $[M + H]^+$ ions of $[D_9]$-N^2-*n*Bu-dG and N^2-heptynyl-dG. 138

Figure 5-1. A schematic diagram of single-base resolution sequencing of N^2-heptynyl-dG in genomic DNA. 147

List of Tables

Table 3-1. Oligonucleotides used in this study.. 102

Table 4-1. Oligonucleotides used in this study.. 139

Chapter 1

General Introduction

Genetic information is faithfully passed down from parent cells to their offspring through the process of cell division. Variations in DNA can arise from alterations occurring in DNA repair, chromosome duplication, transmission, or recombination, creating a natural source of genetic diversity. While genome instability can also occur in bacteria, its implications are more significant in multicellular organisms. Genome instability plays a central role in the development of cancer, and in humans, it also contributes to certain neurodegenerative diseases like amyotrophic lateral sclerosis and myotonic dystrophy (1-4). Based on the specific types of genetic abnormalities observed, scientists have categorized genome instability into three main classes: chromosomal instability (CIN), microsatellite instability (MIN), and nucleotide-associated instability (NIN) (5, 6).

DNA Damage and its biological consequences

NIN is characterized by an increased frequency of deletions, or substitutions, primarily driven by the incorporation of incorrect nucleotides during replication or failures to repair modified bases. Living cells are constantly exposed to a wide range of agents, both endogenous and exogenous, which lead to the occurrence of DNA damage, and ultimately NIN. Endogenous sources of DNA instability involve processes like cellular metabolism, which can cause DNA oxidation, and dysregulated non-enzymatic DNA methylations (7, 8). Conversely, exogenous sources such as ionizing radiation, environmental carcinogens, and the use of chemotherapeutic drugs can induce various types of DNA lesions, including alkylation, oxidation, intra- and inter-strand crosslinks, DNA-protein crosslinks, depurination, deamination, and single-strand breaks (9).

Most of the N^2-alkylguanine lesions are formed from exposure to known mutagens or carcinogens. The N^2-methyl-dG adduct, the smallest alkyl adduct in the series is formed by the reaction N^2-amino group of guanine with formaldehyde or endogenous methylating agents (10, 11). Classified by the International Agency for Research on Cancer (IARC) as a human carcinogen, formaldehyde is an ubiquitous pollutant in vehicle exhaust and cigarette smoke and a common endogenous metabolic byproduct (12). Another well-established biomarker, the N^2-ethyl-dG DNA adduct, is indicative of exposure to acetaldehyde (13, 14). This substance, classified by IARC as an animal carcinogen and a potential human carcinogen, is present as both an exogenous pollutant in cigarette smoke and an endogenous product of ethanol metabolism.

Methylglyoxal (MG) is one of the reactive carbonyl species which can be produced endogenously in all cells and all organisms from the nonenzymatic fragmentation of triose phosphates. In this regard, subjecting human red blood cells to increasing glucose concentrations in vitro could lead to heightened levels of intracellular MG. This elevated MG concentration has also been identified in the kidney (both cortex and medulla), lens, and bloodstream of diabetic rats induced with streptozotocin, as well as in blood samples collected from diabetic patients (15-17). When exposed to MG under physiological conditions and temperatures, DNA predominantly forms N^2-(1-carboxyethyl)-2′-deoxyguanosine (N^2-CEdG), a major stable adduct (18).

Polycyclic aromatic hydrocarbons (PAHs) are widespread environmental pollutants with carcinogenic properties that can be found across various media such as air, water, soil, and sediment (19-21). These compounds are also present at high concentrations in sources like tobacco smoke, foods, and occupational environments (22). PAHs are known to induce the development of malignant tumors in rodents, affecting organs such as the mammary gland, lung, bladder, and skin (23-25). Additionally, there is evidence suggesting that the consumption of

PAHs through food increases the risk of conditions like colorectal adenoma and pancreatic cancer (22). Benzo[*a*]pyrene (BaP) serves as a representative compound for investigating the carcinogenic effects of PAHs. The carcinogenic PAHs, including BaP, necessitate metabolic activation to exhibit mutagenic, teratogenic, and carcinogenic effects (26). BaP undergoes metabolic transformation to produce its ultimate metabolite, trans-7,8-dihydroxy-anti-9,10-epoxy-7,8,9,10-tetrahydrobenzo[*a*]pyrene (BPDE). BPDE is capable of binding with cellular DNA, leading to the formation of BPDE-DNA adducts, with BPDE-dG being a predominant DNA adduct (27).

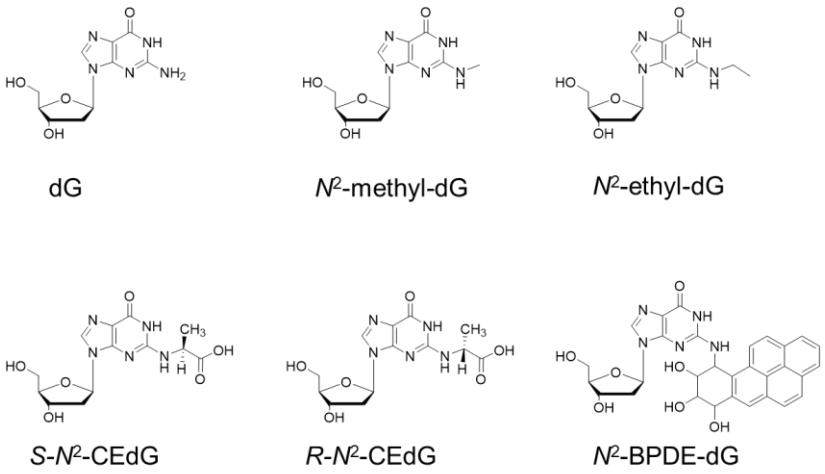

Figure 1-1. Structures of 2′-deoxyguanosine (dG) and N^2-alkylguanine lesions.

In accordance with the specific sites of alkylation, diverse alkyl lesions generated by alkylating agents have the potential to undermine genome integrity. This can manifest through the initiation of mutagenic events and the impediment of critical biological functions, such as DNA replication and transcription. The mutagenic and cytotoxic consequences arising from alkyl lesions hold the potential to culminate in conditions such as cancer and the aging process over time (28-30).

For the readily formed N^2-alkyldG lesions, no mutagenic properties were observed, as the other unreacted hydrogen still maintained the ability to form a hydrogen bond with the correct nucleobase, allowing this type of lesion to be bypassed with high fidelity. Moreover, the cytotoxicity of N^2-alkyldG has also been observed in the transcription process (31). Other studies suggested these alkylated lesions could impede replication by some DNA polymerases, and alter the hydrogen-bonding pattern in duplex DNA (32).

R-loop structures

An R-loop is a three-stranded nucleic acid structure that consists of a DNA:RNA hybrid and a displaced strand of DNA (33). R-loops occur frequently in genomes and have notable physiological significance. They exert crucial control over gene expression, DNA replication, as well as DNA and histone modifications. Numerous investigations have unveiled the pivotal involvement of R-loops in fundamental biological mechanisms across diverse organisms. Interestingly, despite their pivotal positive roles in facilitating crucial biological processes, they can also paradoxically lead to DNA damage and disruptions in genome stability (34, 35).

R-loop structures were shown to occur preferentially near transcriptionally active genes based on genome-wide mapping assay results (36-38). In particular, they tend to form in promoter regions (39). During the transcription process, R-loops frequently form when a nascent transcript rebinds with the template DNA strand (40). There are various factors that can increase the propensity of R-loop formation. First, pausing of RNAPII at transcriptional start sites is correlated with R-loop formation in human cells (41). Head-on collision of the transcription and replication machineries was reported to induce R-loop formation (42). Moreover, R-loop formation is more efficient when a G-rich transcript is produced from a C-rich template. G-quadruplex (G4) structures that form on the displaced strand of DNA can also further stabilize the

R-loop structures (43). In addition, ssDNA nick downstream of the promoter and ionizing-radiation induced DNA damage were documented to be strong, transcription-dependent initiators of transient R-loop formation. This is likely due to the increased displacement potential of the non-template strand, facilitating the initiation of R-loop structures (44).

In general, R-loops may interfere with DNA replication, repair, and transcription, thus compromising genome integrity and function (40, 45). Scheduled formation of R-loop structures plays various physiological roles in regulating genome dynamics such as immunoglobulin (Ig) class-switch recombination (CSR) or CRISPR-Cas9 activity, in which the guide-RNA forms a DNA-RNA hybrid as an intermediate to identify the target for Cas9-mediated cleavage or initiation of DNA replication in mitochondrial DNA (46, 47). In addition, increasing lines of evidence indicates that R-loops play roles in modulating gene expression and chromatin state (36, 48). However, unscheduled R-loop formation leads to adverse effects. In both *E. coli* and human cells, the accumulation of aberrant or excessive R-loops has been demonstrated to contribute to chromosomal instability (CIN), aligning with the concept of genomic instability (46). Exposed ssDNA in R-loop structures can be cleaved by different endonucleases leading to DNA strand breaks and/or mutagenic events; ssDNA can also adopt harmful secondary structures, including G-quadruplexes and hairpins. R-loop accumulation, RNA polymerase stalling, and/or R-loop-driven chromatin condensation can cause transcription block/slow down, replication stress and double-strand breaks (DSBs) (49, 50).

R-loops are clearly emerging to have a central role in cell biology, not only for their physiological roles but also for their pathological implications. An altered R-loop homeostasis has been documented in several diseases, including neurological disorders and cancer (49). The key difference with cancer is that neurodegenerative disorders affect non-dividing neuronal cells; therefore, R-loops accumulated in neuronal cells would not compromise DNA replication. In

cases like Friedreich's ataxia patient cell lines, which carry GAA repeat expansions, R-loops form at such trinucleotide repeats and trigger transcriptional silencing of the genes associated with the disease (51). On the other hand, an overabundance of R-loops at promoters and terminators in BRCA1 mutant cells seems to be directly implicated in tumorigenesis (52-54).

Cells have developed backup mechanisms to resolve R loops and avoid their accumulation and potential threat to transcription and genome integrity. Probably the most relevant and well-known factor with this function is RNase H, a conserved ribonuclease from bacteria to humans which exhibits a specific capacity for degrading the RNA component within RNA: DNA hybrid structures (55). In principle, both forms of RNase H, namely RNase H1 and RNase H2, possess the capability to disentangle these hybrids. Depletion of or inactivation for RNase H1 led to elevated R-loop accumulation (56, 57). Moreover, RNase H1 overexpression efficiently suppresses specific phenotypes associated with R-loops or eliminates in vivo the DNA-RNA hybrid signal detected by the S9.6 antibody. The latter form, RNase H2, also plays a role in ribonucleotide excision repair (RER). However, the key role of RNase H is to remove the RNA primer of the DNA-RNA hybrid in Okazaki fragments. This is a highly efficient back-up mechanism to resolve R-loops, which is supported by the observations in bacteria, yeast, and human cells. However, overexpression of RNase H is an artificial condition and it cannot be excluded that an excess of RNase H could act on certain hybrids that are not necessarily its main natural target.

In recent years, numerous reports have identified an increasing number of RNA-dependent ATPases. Some of these ATPases have demonstrated ability to unwind DNA-RNA structures *in vitro*, and when they were depleted from cells, it results in an accumulation of R-loops (58-62). These results suggest that DNA-RNA helicases constitute a second type of factors to remove R loops. The idea of RNA helicases participating in the resolution of R loops holds conceptual

appeal due to the fundamental contrast with RNases H. Unlike RNases H, helicases would resolve the hybrid without the costly action of degrading the nascent RNA (49).

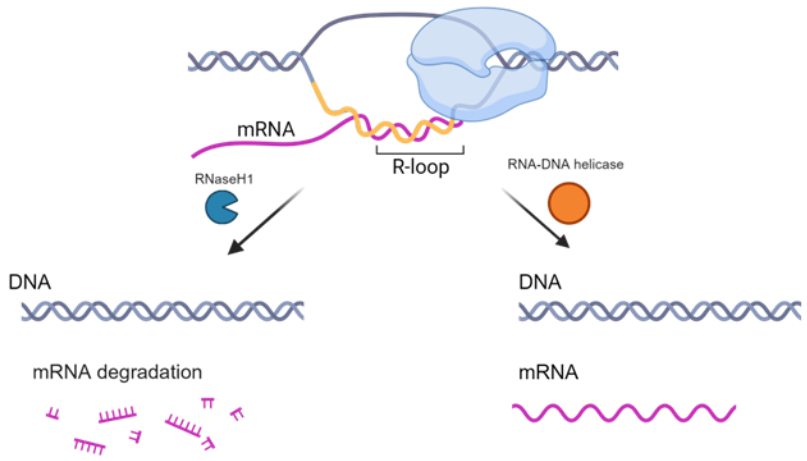

Figure 1-2. R-loop removal is achieved by RNase H enzymes that degrade the RNA moiety of the hybrid and by DNA-RNA helicases that unwind the hybrid.

Translesion Synthesis and DNA polymerases

Although most DNA lesions can be removed by cellular DNA repair pathways, a fraction of DNA lesions can elude detection and endure into the S-phase of the cell cycle. However, replicative DNA polymerases possess highly constrained active sites, making it challenging for them to accommodate damage sites (63). As a result, upon encountering a DNA lesion, the replication machinery becomes incapable of incorporating the correct nucleotide opposite the damaged site, leading to fork stalling and collapse. The incapacity to recommence the replication process subsequently leads to cell cycle arrest and eventual cell death.

In response to this scenario, the translesion synthesis (TLS) mechanism is activated. Translesion synthesis (TLS) stands as a prominent DNA damage tolerance mechanism wherein a

cohort of specialized polymerases, denoted as TLS polymerases, assume the role of the conventional replicative polymerases to facilitate replication across sites harboring lesions. These TLS polymerases possess the capability to introduce nucleotides, utilizing the damaged DNA sites as templates, and are ubiquitously distributed across all three domains of life (64-66). Although TLS has been extensively studied in the recent couple of decades, its molecular mechanism remains incompletely elucidated. In this context, two models have been posited to expound upon the manner in which TLS polymerases effectuate lesion bypass: the polymerase-switching model and the gap-filling model (Figure 1.3) (67). Within the polymerase-switching model, when a replicative polymerase encounters an impediment at the lesion site, a transition occurs via protein-protein interactions, thereby resulting in the displacement of the replicative polymerase by the TLS polymerase. Following the relatively precise bypass of the lesion, the replicative polymerase is subsequently reinstated to resume its role in accurate DNA synthesis. (68-70). Conversely, in the gap-filling model, TLS transpires external to the replication forks. In this scenario, the replication machinery continues its progress, albeit leaving behind a single-strand DNA (ssDNA) gap opposite the DNA lesion. Subsequently, the TLS polymerase is recruited to the site of this gap and proceeds to fill it (71).

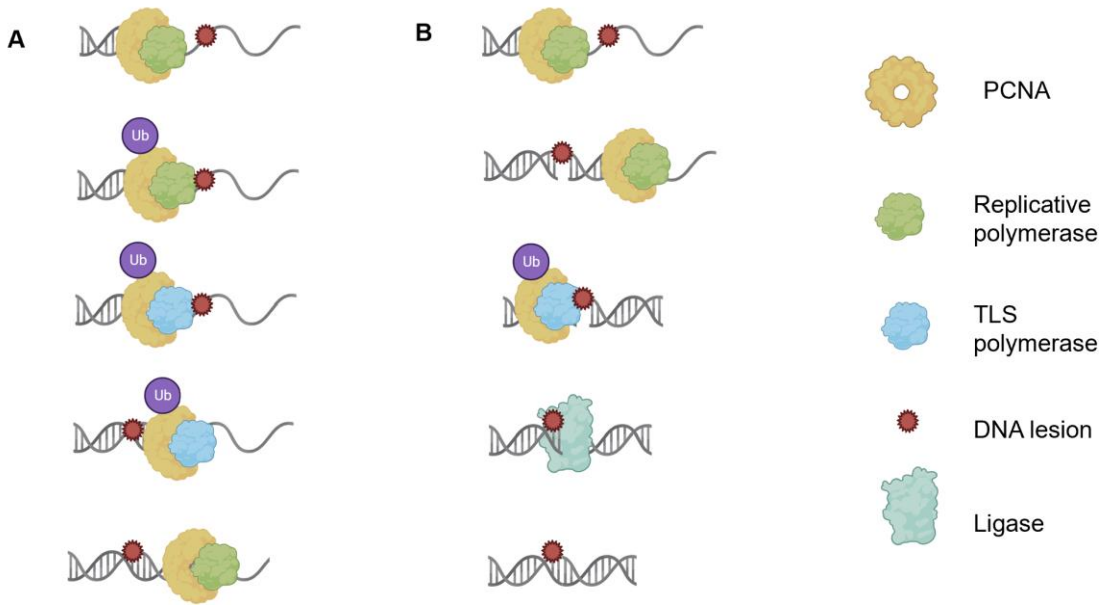

Figure 1-3. Two models for lesion bypass by TLS Pols, A) the polymerase-switching model and B) the gap-filling model.

TLS is performed by a relatively new category of DNA polymerases. Polymerase η (Pol η), Polymerase ι (Pol ι), Polymerase κ (Pol κ), and REV1 in the Y-family and Pol ζ in the B-family are responsible for most TLS in mammalian cells. Moreover, Y-family polymerases Pol IV and Pol V, and B-family polymerase Pol II have been founded in *E. coli* (72-76). These TLS polymerases have active sites that are larger and more open than those of the high-fidelity replicative DNA polymerases (Pol α, δ, and ε), allowing accommodation of and synthesis past DNA templates with large, helix-distorting lesions (77). Many TLS polymerases are conserved throughout all kingdoms of life, and were reported to be equipped with high efficiency and fidelity in bypassing damaged DNA bases. For example, Pol η is reported to bypass thymine-thymine cyclobutane pyrimidine dimers (TT-CPDs), 8-oxodG as well as the 1,2-intrastrand d(GpG)-cisplatin crosslink adduct (78-80). Pol κ has shown its adeptness in bypassing the minor-groove N^2-alkyldG lesions, such as N^2-CEdG, and N^2-furfuryl-dG (81, 82).

In addition to their primary role in lesion bypass, human TLS polymerases have been documented to exhibit non-TLS-related biochemical activities. For instance, human DNA polymerase η has been observed to incorporate ribonucleoside triphosphates (rNTPs) opposite both undamaged and damaged nucleosides, such as 8-oxo-dG and TT-CPD, However, the rates of incorporation for rNTPs are significantly lower than those observed for the corresponding deoxyribonucleoside triphosphates (dNTPs) (83, 84). Additionally, human REV1 displays a preferential capacity for catalyzing rCTP incorporation, albeit with an efficiency approximately 280-fold lower than that seen with dCTP insertion (83). Recent investigations have unveiled novel roles for TLS polymerases in the resolution of non-B-form DNA structures. It has been documented that the catalytic core of REV1 possesses the capability to directly destabilize G4 structures and prevent their refolding (85, 86). Furthermore, both Pol η and Pol κ have been identified as pivotal contributors to the maintenance of replication fork progression past naturally occurring non-B-form DNA structures, including G4 DNA (87-89).

Methods

In vivo replication and transcription experiments

To evaluate the impact of DNA lesions on the efficiency and accuracy of replication or transcription, and to investigate the contributions of TLS polymerases in bypassing these lesion sites, several assays involving polyacrylamide gel electrophoresis (PAGE) and mass spectrometry have been developed.

1. *In vivo* Replication Experiment in *E.* coli cells

Originally devised by Delaney and Essigmann (90), the competitive replication and adduct bypass (CRAB) assay is a method employed to evaluate the interference caused by DNA lesions during DNA replication within *E. coli* cells utilizing bacteriophage M13 as a model system (91,

92). In this assay, lesion-containing and lesion-free M13 genomes are first constructed, plus a competitor M13 genome, which serves as an internal standard. Subsequently, the lesion-bearing or lesion-free plasmids, are pre-mixed with the competitor genome and collectively introduced into *E. coli* cells. After undergoing replication within the *E. coli* cells, a specific genomic segment of the resulting M13 progeny, which includes the original lesion site, is selectively amplified using polymerase chain reaction (PCR). The resultant PCR products are then subjected to digestion using appropriate restriction enzymes, and the resulting digestion products are subjected to analytical techniques including liquid chromatography-tandem mass spectrometry (LC-MS/MS) and polyacrylamide gel electrophoresis (PAGE) analyses (Figure 1.4). This assay enables the quantification of bypass efficiency and mutation frequency specific to individual lesions. Moreover, by conducting replication experiments in *E. coli* cells that are either proficient or deficient in specific repair proteins or TLS polymerases, it becomes feasible to elucidate the respective roles of these factors in the repair and translesion synthesis of various DNA lesions.

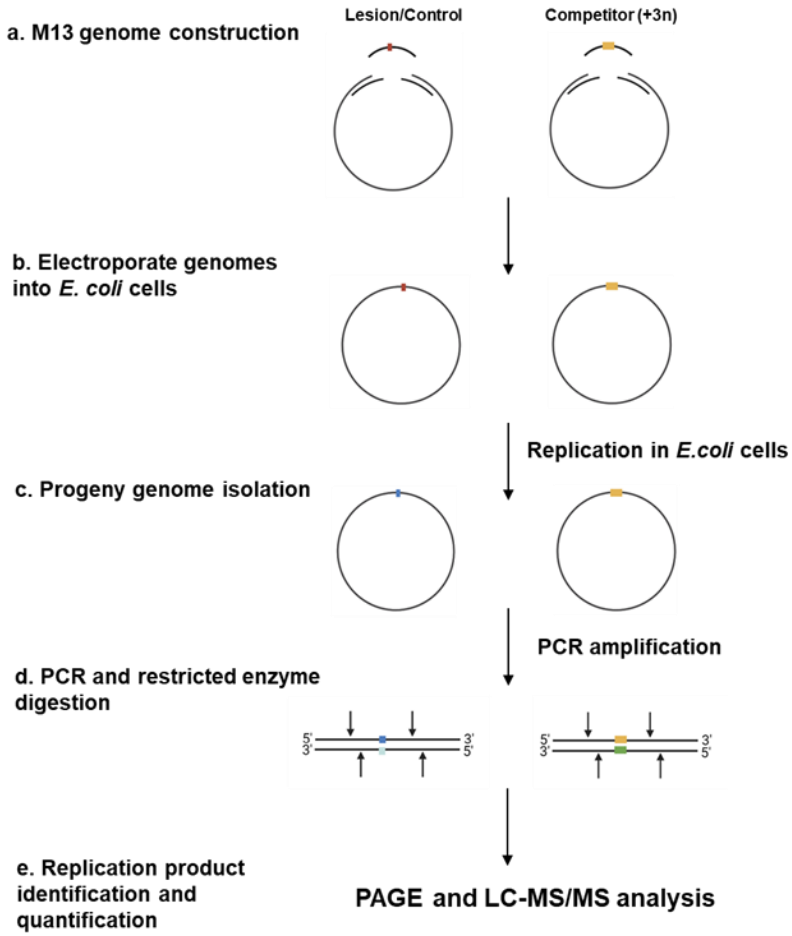

Figure 1-4. A schematic diagram illustrating the experimental procedures of CRAB assay.

2. *In vivo* Transcription Experiment in Mammalian Cells

The objectives of this investigation encompassed the evaluation of how DNA lesions influence transcription. To achieve these aims, we employed a previously established competitive transcription and adduct bypass (CTAB) assay. Initially, we created double-stranded, non-replicative plasmids that contained DNA lesions incorporated at specific sites using a gapped

vector-based approach (93). These lesions were strategically positioned downstream of a CMV promoter to facilitate transcription mediated by human RNA polymerase II (RNAPII).

The production of the gapped vector entails a multi-step process. Initially, a parental vector, specifically the pTGFP-T7-Hha10T plasmid, is subjected to nicking by the enzyme Nt.BstNBI. Subsequently, the resulting nicked plasmid undergoes a removal step involving a 25-mer oligodeoxyribonucleotide (ODN), resulting in the generation of the gapped vector. This gapped vector is then subjected to ligation, wherein it is combined with a 13-mer ODN devoid of lesions, as well as a 12-mer ODN containing the DNA lesion of interest (Figure 1.5). Similar to the assay performed in bacteria, we subsequently premix lesion-containing or lesion-free double-stranded plasmid and transfected into HEK293T cells. After cellular transcription, the runoff transcripts were reverse-transcribed with a gene-specific primer and the resulting cDNA was amplified by PCR. The PCR products were digested with restriction enzymes, the digestion products are subjected to PAGE and LC-MS/MS analyses (Figure 1.6).

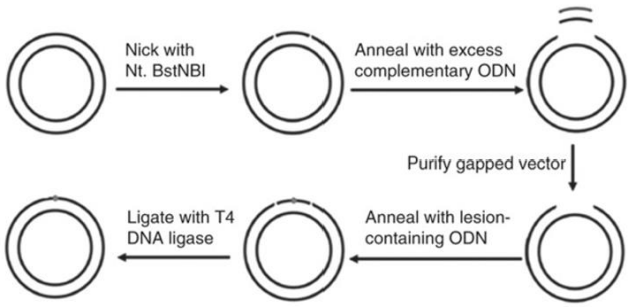

Figure 1-5. A schematic diagram illustrating the procedures for the construction of plasmids harboring a site-specifically incorporated lesion (93).

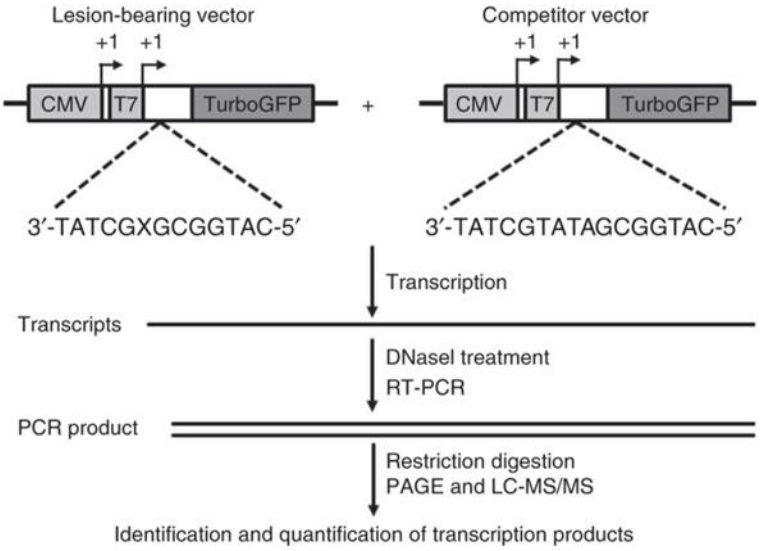

Figure 1-6. A schematic diagram illustrating the CTAB assay system (93).

Interaction Proteomics

Proteomics, a field devoted to investigating the properties, functions, compositions, and structural aspects of proteins, as well as their roles in cellular processes, has achieved noteworthy advancements in recent years (94). Particularly, mass spectrometry-based proteomics has emerged as a pivotal tool in the domains of molecular and cellular biology, as well as the burgeoning field of systems biology. These achievements encompass diverse areas such as the elucidation of protein-protein interactions, the comprehensive mapping of various organelles, the simultaneous delineation of the genome and proteome of the malaria parasite, and the generation of quantitative profiles of proteins across a wide spectrum of species. The potential of mass spectrometry to not only identify but also quantify numerous proteins from intricate biological samples holds considerable promise for applications spanning the realms of biology and pharmacology. It is worth noting that the progression of biological functions and the onset of human diseases are intricately linked to protein-protein interactions (PPIs), which are characterized by factors such as proximity, affinity, and duration.

1. Principles and Instrumentation

Mass spectrometry entails measurements conducted in the gaseous phase on ionized analytes. Fundamentally, a mass spectrometer comprises three core components: an ion source, a mass analyzer tasked with determining the mass-to-charge ratio (*m/z*) of ionized analytes, and a detector responsible for recording ion counts at specific *m/z* values. The electrospray ionization (ESI) source is utilized to ionize analytes from a solution, rendering it compatible with liquid-based separation techniques such as chromatography and electrophoresis. Integrated systems of LC-MS are the preferred choice for analyzing complex samples. The mass analyzer stands as a central and pivotal element in this technology. Within the realm of proteomics, it is characterized by key parameters including sensitivity, resolution, mass accuracy, and its ability to generate information-rich mass spectra of peptide fragments, known as MS/MS. Currently, proteomics research employs four fundamental types of mass analyzers: ion trap, time-of-flight (TOF), quadrupole, and Fourier-transform ion cyclotron resonance (FT-MS), orbitrap analyzers. Furthermore, before peptides undergo analysis via mass spectrometry, it is imperative to fragment them further through a process termed dissociation. Multiple dissociation techniques are available for mass spectrometry instruments, including collision-induced dissociation (CID), electron-capture dissociation (ECD), and electron-transfer dissociation (ETD).

2. Bottom-up proteomics

The application of mass spectrometry (MS)-based proteomics has paved the way for the comprehensive characterization and quantitative analysis of the protein composition within biological specimens, thereby enabling the exploration of intricate molecular interactions associated with distinct pathological conditions. Notably, in the area of cancer research, the integration of diverse proteomic methodologies with genomic analysis has proven instrumental in

elucidating the molecular underpinnings of tumorigenesis and the advancement of more efficacious strategies for combating cancer (95-97).

"Bottom-up" protein analysis pertains to the characterization of proteins through the examination of peptides liberated from the proteins via proteolysis. When bottom-up proteomics is performed on a mixture of proteins it is called shotgun proteomics. Bottom-up proteomics capitalizes on the inherent advantages offered by peptides over intact proteins: peptides are more amenable to separation through reversed-phase liquid chromatography (RPLC), exhibit favorable ionization properties (98), and undergo fragmentation in a more predictable manner. Today, bottom-up approaches using data-dependent acquisition (DDA) workflows are the core technologies in proteomics. A typical bottom-up proteomics workflow (Figure 1.7) consists of several major steps: isolation of the protein mixture from the target biological sample, followed by quantification of protein concentrations within the isolated sample, often utilizing methods such as the Bradford assay. Subsequently, the proteins are fractionated using gel electrophoresis. Following fractionation, proteolytic cleavage of the proteins is carried out, typically utilizing trypsin, which generates peptides featuring basic arginine or lysine residues at their C-termini. Next, a mass spectrometric analysis is conducted on the resultant peptides, followed by a database search aimed at identifying the proteins (99). Peptide identification is accomplished by comparing the tandem mass spectra derived from peptide fragmentation to the theoretical tandem mass spectra generated through *in silico* digestion of a protein database.

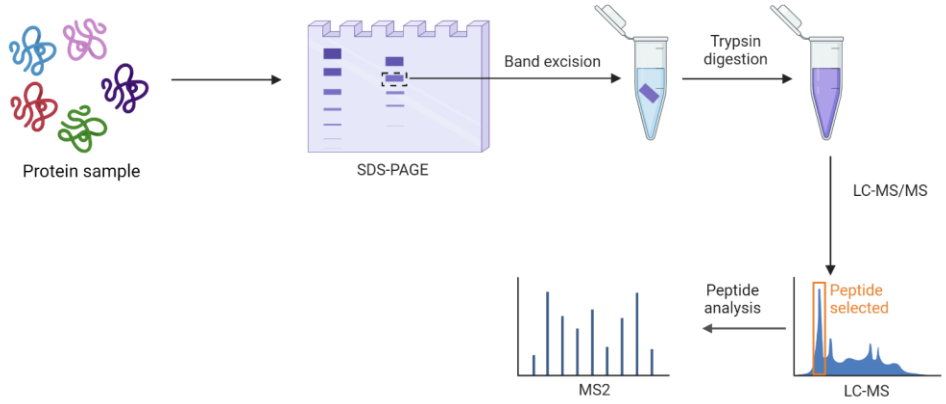

Figure 1-7. Experimental workflows for bottom-up proteomic approaches.

3. Ascorbate peroxidase (APEX) Proximity labeling

Over the past decade, enzyme-catalyzed proximity labeling (PL) has emerged as a novel method, providing an alternative approach for tagging and capturing not only the proteins that directly interact with the protein of interest (POI) but also those in close proximity to the POI. In a PL system, a promiscuous labeling enzyme is genetically fused to the POI or subcellular marker proteins within live cells. One such enzyme, APEX, is a monomeric 28 kDa ascorbate peroxidase. It can be directed to a specific cellular location through fusion with a POI or a subcellular targeting motif. In the presence of hydrogen peroxide (H_2O_2), APEX catalyzes the oxidation of an exogenous biotin-phenol (BP) to phenoxyl radicals. These radicals can subsequently covalently bind to electron-rich amino acids, such as tyrosine, in proteins located in close proximity to the APEX active site. Notably, the efficiency of biotinylation significantly diminishes within a few nanometers of the APEX active site, resulting in a confined labeling radius (< 20 nm) achieved within a short time frame (100, 101). The biotinylated endogenous proteins are subsequently enriched using streptavidin-coated beads and can be identified through mass spectrometry.

APEX offers the advantage of rapid labeling, capturing nearby proteins within minutes, as opposed to hours required by other methods like BioID (102). However, the APEX labeling approach involves the overexpression of APEX-fusion proteins, which may introduce potential artifacts when identifying interacting proteins. Additionally, the APEX method relies on the biotinylation of amino acids situated near the target POI. Since only a limited number of electron-rich amino acids can be biotinylated, it is possible that certain interaction partners may evade labeling and subsequent detection via LC-MS/MS. Furthermore, biotin-phenol reagents can be cytotoxic, making APEX less suitable for organoid or in vivo studies (103).

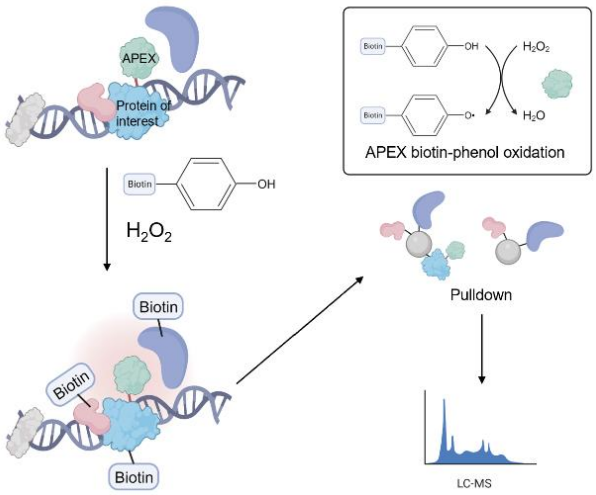

Figure 1-8. General workflow of proximity labeling followed by mass spectrometry with peroxidase.

4. Affinity purification

Endogenous proteins can serve as effective bait molecules for the isolation of protein complexes, provided that specific antibodies or other reagents exist for their selective purification along with their associated binding partners. Regrettably, comprehensive collections of antibodies are not presently available, and many existing antibodies exhibit limitations such as poor

immunoprecipitation efficiency or insufficient specificity. As a more versatile alternative, a strategy involves the "tagging" of proteins of interest with a recognizable sequence to enable their affinity-based purification using specific antibodies raised against the tag. In affinity purification mass spectrometry experiments, epitope tagging is frequently employed. In this approach, short peptide or protein tags, such as FLAG-, Strep-Tag, or c-myc (104) are genetically fused to the POI. This fusion can occur either within an exogenous expression construct or within the native gene locus to facilitate the expression of the tagged protein at levels that closely resemble its physiological counterpart. Ideally, the tagged construct is expressed under the control of the promoter governing the expression of the untagged native protein.

To gain insights into the functionality of proteins within intricate biological systems, researchers require tools for the visualization of endogenous proteins. The process of labeling endogenous proteins offers distinct advantages when compared to alternative methods. By tagging endogenous proteins, potential issues related to the mis-localization of proteins due to overexpression are circumvented. Additionally, this approach preserves the inherent transcriptional and posttranscriptional regulation of the gene. The advent of CRISPR/Cas9 technology has revolutionized the field of gene editing, simplifying and accelerating genome editing processes while also reducing costs (105, 106). In this context, we have developed an interface that bridges the gap between genome editing and proteomics. This interface allows us to isolate native protein complexes in their natural genomic contexts (107-110). Specifically, we utilized CRISPR-Cas9 to introduce three consecutive Flag epitope tags at the C-terminus of endogenous proteins of interest within HEK293T cells. Subsequently, we enriched these tagged proteins using anti-Flag M2 beads and conducted LC-MS/MS analysis for further investigation (Figure 1.9).

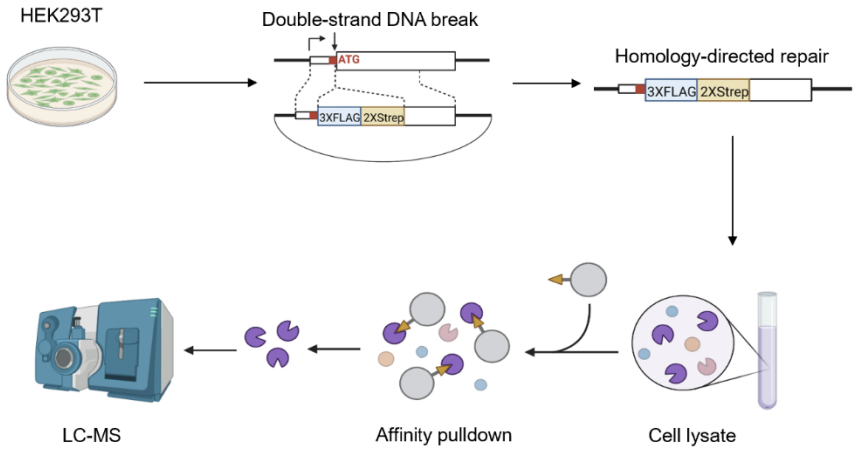

Figure 1-9. General workflow of CRISPR-KI followed by mass spectrometry with peroxidase.

Next-generation Sequencing (NGS)

By the year 1986, a pivotal milestone was achieved with the development of the first automated DNA sequencing method, marking the onset of a remarkable period characterized by the advancement and refinement of sequencing platforms (Figure 1.10) (111). This method, famously known as the chain-termination technique or Sanger sequencing, employs a DNA sequence of interest as a template for PCR, during which dideoxyribonucleoside triphosphates (ddNTPs) are incorporated into the growing DNA strand (112). In automated Sanger sequencing, all four ddNTPs are concurrently present in a single reaction, with each dNTP carrying a distinct fluorescent label. When DNA polymerase integrates a ddNTP, it effectively terminates the extension process, resulting in the generation of multiple copies of the DNA sequence, encompassing all possible lengths within the amplified fragment. Subsequently, these chain-terminated oligonucleotides are separated by size. In earlier methods, gel electrophoresis was employed for this purpose, whereas later automated capillary sequencers used capillary tubes to achieve the size separation. Ultimately, the DNA sequence is determined based on the resulting pattern of fluorescently labeled fragments.

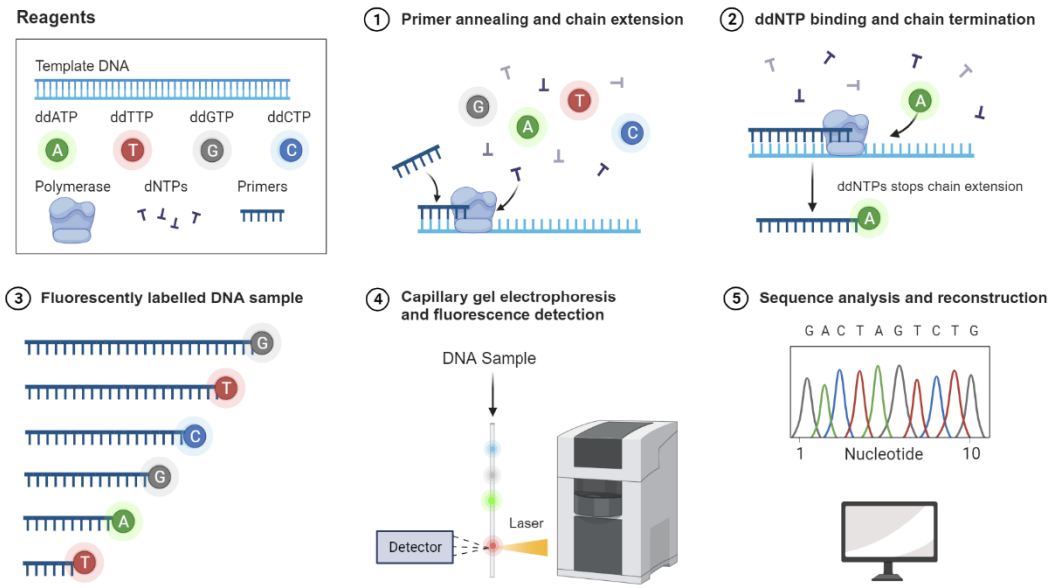

Figure 1-10. A schematic workflow of Sanger sequencing.

In the year 2005, a significant milestone was achieved with the introduction of the first commercially available next-generation sequencing (NGS) platform. This technology revolutionized DNA sequencing by enabling the simultaneous amplification of millions of copies of a specific DNA fragment in a massively parallel manner, in stark contrast to Sanger sequencing (Figure 1.11). While there are fundamental similarities in the underlying principles between Sanger sequencing and NGS, the key distinction lies in their sequencing throughput and capacity. In NGS, the genetic material (DNA or RNA) is fragmented first. Subsequently, oligonucleotides with known sequences are attached to these fragments through a process called adapter ligation. This modification allows the fragments to interact effectively with the chosen sequencing system, wherein the bases of each fragment are identified based on the signals they emit (113). The principal divergence between Sanger sequencing and NGS resides in their sequencing volumes. NGS empowers the concurrent processing of millions of reactions in parallel, resulting in significantly enhanced throughput, heightened sensitivity, expedited sequencing speed, and

reduced overall cost. Consequently, numerous genome sequencing projects that would have previously spanned years when utilizing Sanger sequencing methods can now be completed within a matter of hours using NGS technology.

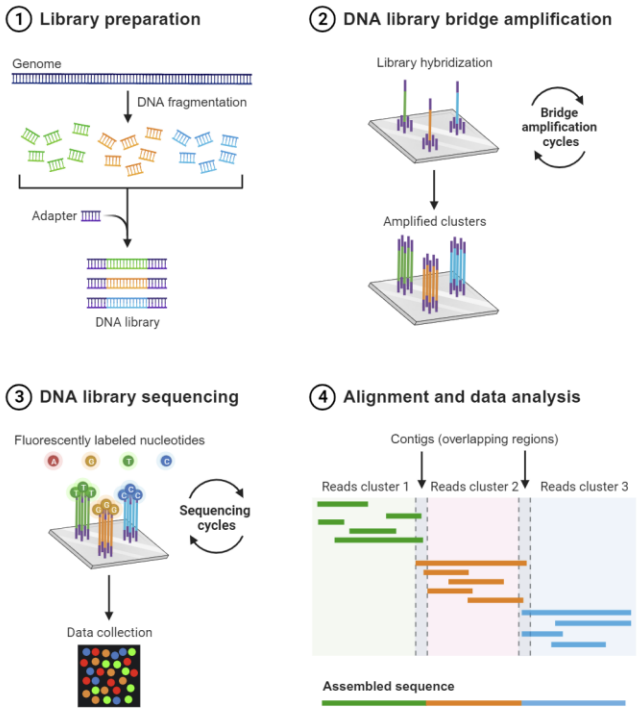

Figure 1-11. A schematic diagram of NGS.

1. Chromatin immunoprecipitation sequencing (ChIP-Seq)

Chromatin immunoprecipitation (ChIP) assays, when combined with sequencing technology, give rise to a powerful method known as ChIP sequencing (ChIP-Seq), which is instrumental for the comprehensive identification of genome-wide DNA binding sites for various proteins, including transcription factors (Figure 1.12). In the course of ChIP protocols, the DNA-binding protein is crosslinked to DNA within live cells through formaldehyde treatment. Subsequently, the chromatin is fragmented into small segments, typically within the range of 200 to 600 base pairs, via sonication (114). DNA associated with a protein of interest is first immunoprecipitated

using a specific antibody. The DNA that is bound to the protein is subsequently co-precipitated, purified, and subjected to sequencing. The integration of next-generation sequencing (NGS) into ChIP experiments has provided valuable insights into the mechanisms of gene regulation implicated in diverse diseases and biological pathways, including developmental processes and cancer progression. ChIP-Seq offers the capability to thoroughly investigate protein-nucleic acid interactions on a genome-wide scale.

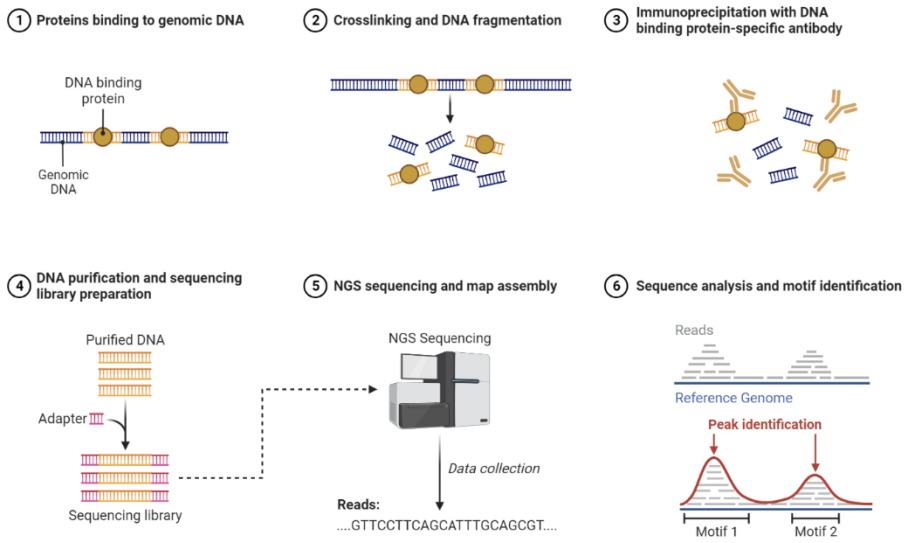

Figure 1-12. An outline of ChIP-seq workflow. Immunoprecipitated DNA can be analyzed by next-generation sequencing, to provide a genome-wide overview of target enrichment or a location-specific analysis.

2. R-loop sequencing

Accurate mapping of R-loop locations is important for comprehending their functional roles. The genome-wide detection of R-loops predominantly hinges on the enrichment of RNA-DNA duplexes, achieved through the utilization of the S9.6 monoclonal antibody (36, 114-116), or catalytically inactive RNase H1 (41, 117), followed by high-throughput sequencing. While

these methods have proven effective in elucidating R-loop functions, immunoprecipitation-based approaches typically necessitate a large number of cells, limiting their applicability to biological processes with constrained input materials. Furthermore, it is noteworthy that the S9.6 antibody and the inactive RNase H1 exhibit distinct preferences for different R-loop sequences, yielding variable outcomes across different methodologies (45). For instance, in DNA:RNA immunoprecipitation sequencing (DRIP-seq), R-loop signals are detected throughout the entirety of gene-coding regions. Conversely, bis-DRIP-seq (118) and R-ChIP (112) predominantly identify R-loops at promoter regions. Consequently, there exists a compelling need for an R-loop mapping approach that does not rely on the S9.6 antibody or RNase H1 and is amenable to applications in live cells, particularly when dealing with limited input materials.

A recent development in the field involves the utilization of kethoxal-assisted single-stranded DNA (ssDNA) sequencing, abbreviated as KAS-seq (Figure 1.13), for the comprehensive genome-wide mapping of single-stranded DNA regions (119). N_3-kethoxal, a chemical reagent, selectively labels the N1 and N2 positions of guanine residues, rendering it highly specific for ssDNA and incapable of reacting with double-stranded DNA (dsDNA) (120). Within transcription bubbles, N_3-kethoxal interacts with both DNA strands when nascent RNA is not engaged in base pairing with the DNA template strand. However, in the context of R-loops, it exclusively labels the exposed DNA strand, excluding the other strand involved in forming the RNA-DNA duplex (121). The experimental procedure entails labeling live cells with N_3-kethoxal, followed by the isolation of genomic DNA (gDNA). Subsequently, 'click' chemistry is employed to introduce biotin moieties onto the gDNA. The resulting DNA fragments, enriched through biotin-streptavidin interactions, are then subjected to fragmentation and employed in the construction of sequencing libraries (119).

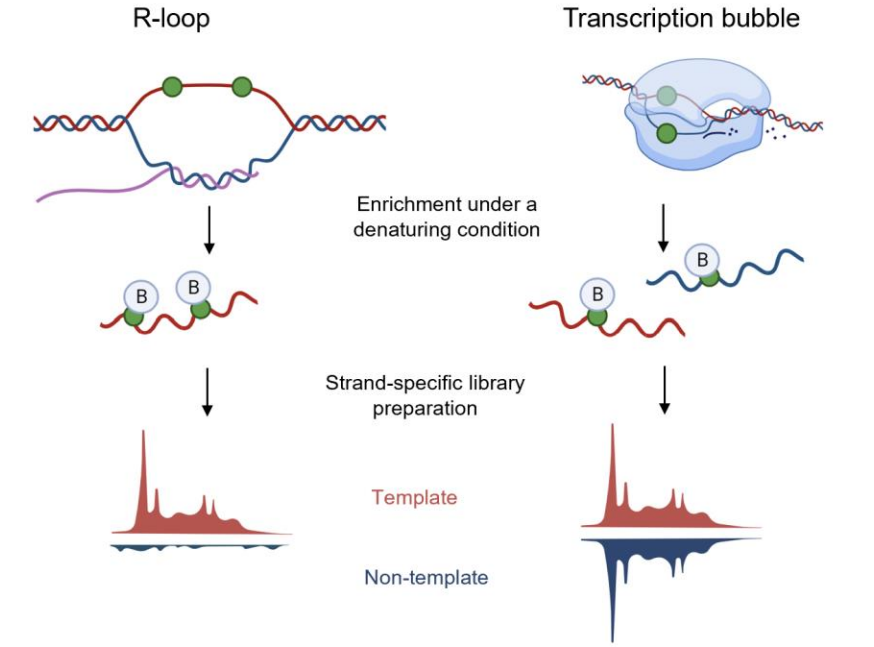

Figure 1-13. A schematic illustration of the principle that spKAS-seq distinguishes R-loops from transcription bubbles. The RNA-DNA hybrid duplex in R-loops blocks one DNA strand from N_3-kethoxal (green dots) labeling, resulting in an unbalanced spKAS-seq read density on two DNA strands. Pull-down was performed under a denaturing condition to ensure capturing of only the N_3-kethoxal–modified strand.

Scope of this Dissertation

The focus of this dissertation is placed on how alkylated DNA lesions and R-loop structure trigger genome instability and biological functional studies of TLS polymerases.

In Chapter 2, we synthesized ODNs harboring N^2-alkyldG lesions at a defined site and ligated them into single-stranded M13 vectors. By performing the CRAB assay in wild-type and TLS polymerase deficient *E. coli* cells, we investigated the impact of size, shape of these alkyl

groups on DNA replication in *E. coli* cells and revealed the roles of each TLS polymerase in supporting the replicative bypass of those lesions.

In Chapter 3, we moved on to the interaction proteome study of human Pol κ to explore new functions of Pol κ. We used two independent approaches (proximity labeling and affinity pull-down) followed by liquid chromatography-tandem mass spectrometry (LC-MS/MS) analysis to profile Pol κ-interacting proteins at the proteome-wide level. Moreover, by performing ChIP-seq and immunofluorescence assays, we investigated the non-canonical function of Pol κ in resolving R-loop structure at the genome-wide scale.

In Chapter 4, we incubated HEK 293T cells with BPDE and N^2-*n*BudG adduct and performed immunofluorescences assay to demonstrate N^2-dG adducts induce R-loop formation in genome. By conducting ChIP-qPCR and CTAB assay based on lesion-bearing vector, we investigated the N^2-dG lesions induced R-loop structure would impede transcription process. Furthermore, after incorporating synthetic N^2-alkyl-dG nucleoside into genomic DNA, with the R-loop sequencing, we unveiled that N^2-alkyl-dG lesions induce R-loop formation in chromatin.

References

1. L. R. Ferguson, H. Chen, A. R. Collins *et al.*, Genomic instability in human cancer: Molecular insights and opportunities for therapeutic attack and prevention through diet and nutrition. *Semin. Cancer Biol.* **35**, S5-S24 (2015).
2. C. E. Cunningham, M. J. MacAuley, G. Yadav *et al.*, Targeting the CINful genome: Strategies to overcome tumor heterogeneity. *Prog Biophys Mol Biol* **147**, 77-91 (2019).
3. S. Negrini, V. G. Gorgoulis, T. D. Halazonetis, Genomic instability--an evolving hallmark of cancer. *Nat Rev Mol Cell Biol* **11**, 220-228 (2010).
4. L. J. Marnett, J. P. Plastaras, Endogenous DNA damage and mutation. *Trends Genet* **17**, 214-221 (2001).
5. R. De Bont, N. van Larebeke, Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis* **19**, 169-185 (2004).
6. M. Katerji, P. J. Duerksen-Hughes, DNA damage in cancer development: special implications in viral oncogenesis. *Am J Cancer Res* **11**, 3956-3979 (2021).
7. K. Lu, S. Craft, J. Nakamura, B. C. Moeller, J. A. Swenberg, Use of LC-MS/MS and stable isotopes to differentiate hydroxymethyl and methyl DNA adducts from formaldehyde and nitrosodimethylamine. *Chem. Res. Toxicol.* **25**, 664-675 (2012).
8. J. D. Mcghee, P. H. V. Hippel, Formaldehyde as a Probe of DNA Structure .2. Reaction with Endocyclic Imino Groups of DNA Bases. *Biochemistry-us.* **14**, 1297-1303 (1975).
9. N. Shrivastav, B. I. Fedeles, D. Y. Li *et al.*, A Chemical Genetics Analysis of the Roles of Bypass Polymerase DinB and DNA Repair Protein AlkB in Processing N-2-Alkylguanine Lesions In Vivo. *Plos One* **9** (2014).
10. P. J. Brooks, J. A. Theruvathu, DNA adducts from acetaldehyde: implications for alcohol-related carcinogenesis. *Alcohol.* **35**, 187-193 (2005).
11. M. Y. Wang, N. X. Yu, L. Chen *et al.*, Identification of an acetaldehyde adduct in human liver DNA and quantitation as N-2-ethyldeoxyguanosine. *Chem. Res. Toxicol.* **19**, 319-324 (2006).
12. P. J. Thornalley, The Glyoxalase System - New Developments Towards Functional-Characterization of a Metabolic Pathway Fundamental to Biological Life. *Biochem. J.* **269**, 1-11 (1990).
13. L. A. Fothergill-Gilmore, P. A. Michels, Evolution of glycolysis. *Prog Biophys Mol Biol* **59**, 105-235 (1993).
14. P. J. Thornalley, Pharmacology of methylglyoxal: Formation, modification of proteins and nucleic acids, and enzymatic detoxification - A role in pathogenesis and antiproliferative chemotherapy. *Gen. Pharmacol-vasc. S* **27**, 565-573 (1996).
15. M. Frischmann, C. Bidmon, J. Angerer, M. Pischetsrieder, Identification of DNA adducts of methylglyoxal. *Chem. Res. Toxicol.* **18**, 1586-1592 (2005).
16. D. H. Phillips, Fifty years of benzo(a)pyrene. *Nature* **303**, 468-472 (1983).
17. S. S. Hecht, Tobacco smoke carcinogens and lung cancer. *J Natl Cancer Inst* **91**, 1194-1210 (1999).
18. C. E. Bostrom, P. Gerde, A. Hanberg *et al.*, Cancer risk assessment, indicators, and guidelines for polycyclic aromatic hydrocarbons in the ambient air. *Environ Health Perspect* **110 Suppl 3**, 451-488 (2002).
19. B. Bukowska, P. Sicinska, Influence of Benzo(a)pyrene on Different Epigenetic Processes. *Int J Mol Sci* **22** (2021).

20. D. Li, P. F. Firozi, L. E. Wang *et al.*, Sensitivity to DNA damage induced by benzo(a)pyrene diol epoxide and risk of lung cancer: a case-control analysis. *Cancer Res.* **61**, 1445-1450 (2001).
21. E. Kriek, M. Rojas, K. Alexandrov, H. Bartsch, Polycyclic aromatic hydrocarbon-DNA adducts in humans: relevance as biomarkers for exposure and cancer risk. *Mutat Res* **400**, 215-231 (1998).
22. E. Gyorffy, L. Anna, Z. Gyori *et al.*, DNA adducts in tumour, normal peripheral lung and bronchus, and peripheral blood lymphocytes from smoking and non-smoking lung cancer patients: correlations between tissues and detection by 32P-postlabelling and immunoassay. *Carcinogenesis* **25**, 1201-1209 (2004).
23. B. Bukowska, K. Mokra, J. Michalowicz, Benzo[a]pyrene-Environmental Occurrence, Human Exposure, and Mechanisms of Toxicity. *Int J Mol Sci* **23** (2022).
24. A. L. Piberger, C. T. Kruger, B. M. Strauch, B. Schneider, A. Hartwig, BPDE-induced genotoxicity: relationship between DNA adducts, mutagenicity in the in vitro PIG-A assay, and the transcriptional response to DNA damage in TK6 cells. *Arch. Toxicol.* **92**, 541-551 (2018).
25. B. Singer, Invivo Formation and Persistence of Modified Nucleosides Resulting from Alkylating-Agents. *Environ. Health Persp* **62**, 41-48 (1985).
26. G. D. Jones, R. C. Le Pla, P. B. Farmer, Phosphotriester adducts (PTEs): DNA's overlooked lesion. *Mutagenesis* **25**, 3-16 (2010).
27. B. Ma, P. W. Villalta, A. T. Zarth *et al.*, Comprehensive High-Resolution Mass Spectrometric Analysis of DNA Phosphate Adducts Formed by the Tobacco-Specific Lung Carcinogen 4-(Methylnitrosamino)-1-(3-pyridyl)-1-butanone. *Chem. Res. Toxicol.* **28**, 2151-2159 (2015).
28. D. Fu, J. A. Calvo, L. D. Samson, Balancing repair and tolerance of DNA damage caused by alkylating agents. *Nat Rev Cancer* **12**, 104-120 (2012).
29. F. Drablos, E. Feyzi, P. A. Aas *et al.*, Alkylation damage in DNA and RNA--repair mechanisms and medical significance. *DNA Repair (Amst)* **3**, 1389-1407 (2004).
30. J. Wu, H. Du, L. Li *et al.*, The Impact of Minor-Groove N(2)-Alkyl-2'-deoxyguanosine Lesions on DNA Replication in Human Cells. *ACS Chem. Biol.* **14**, 1708-1716 (2019).
31. Y. A. Hegazy, C. M. Fernando, E. J. Tran, The balancing act of R-loop biology: The good, the bad, and the ugly. *J. Biol. Chem.* **295**, 905-913 (2020).
32. I. M. Okazaki, K. Kinoshita, M. Muramatsu, K. Yoshikawa, T. Honjo, The AID enzyme induces class switch recombination in fibroblasts. *Nature* **416**, 340-345 (2002).
33. C. C. So, A. Martin, DSB structure impacts DNA recombination leading to class switching and chromosomal translocations in human B cells. *PLoS Genet* **15**, e1008101 (2019).
34. P. A. Ginno, P. L. Lott, H. C. Christensen, I. Korf, F. Chedin, R-Loop Formation Is a Distinctive Characteristic of Unmethylated Human CpG Island Promoters. *Mol. Cell* **45**, 814-825 (2012).
35. S. R. Hartono, A. Malapert, P. Legros *et al.*, The Affinity of the S9.6 Antibody for Double-Stranded RNAs Impacts the Accurate Mapping of R-Loops in Fission Yeast. *J. Mol. Biol.* **430**, 272-284 (2018).
36. W. Xu, H. Xu, K. Li *et al.*, The R-loop is a common chromatin feature of the Arabidopsis genome. *Nature Plants* **3**, 704-714 (2017).
37. L. A. Sanz, S. R. Hartono, Y. W. Lim *et al.*, Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals. *Mol. Cell* **63**, 167-178 (2016).

38. A. Aguilera, T. Garcia-Muse, R loops: from transcription byproducts to threats to genome stability. *Mol. Cell* **46**, 115-124 (2012).
39. L. Chen, J. Y. Chen, X. Zhang *et al.*, R-ChIP Using Inactive RNase H Reveals Dynamic Coupling of R-loops with Transcriptional Pausing at Gene Promoters. *Mol. Cell* **68**, 745-+ (2017).
40. A. Helmrich, M. Ballarino, L. Tora, Collisions between Replication and Transcription Complexes Cause Common Fragile Site Instability at the Longest Human Genes. *Mol. Cell* **44**, 966-977 (2011).
41. D. Roy, M. R. Lieber, G Clustering Is Important for the Initiation of Transcription-Induced R-Loops In Vitro, whereas High G Density without Clustering Is Sufficient Thereafter. *Mol. Cell Biol.* **29**, 3124-3133 (2009).
42. D. Roy, Z. Zhang, Z. F. Lu, C. L. Hsieh, M. R. Lieber, Competition between the RNA Transcript and the Nontemplate DNA Strand during R-Loop Formation In Vitro: a Nick Can Serve as a Strong R-Loop Initiation Site. *Mol. Cell Biol.* **30**, 146-159 (2010).
43. M. P. Crossley, M. Bocek, K. A. Cimprich, R-Loops as Cellular Regulators and Genomic Threats. *Mol. Cell* **73**, 398-411 (2019).
44. K. F. Yu, F. Chedin, C. L. Hsieh, T. E. Wilson, M. R. Lieber, R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells. *Nat. Immunol.* **4**, 442-451 (2003).
45. M. Jinek, K. Chylinski, I. Fonfara *et al.*, A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821 (2012).
46. C. Grunseich, I. X. Wang, J. A. Watts *et al.*, Senataxin Mutation Reveals How R-Loops Promote Transcription by Blocking DNA Methylation at Gene Promoters. *Mol. Cell* **69**, 426-+ (2018).
47. T. Garcia-Muse, A. Aguilera, R Loops: From Physiological to Pathological Roles. *Cell* **179**, 604-618 (2019).
48. R. E. Wellinger, F. Prado, A. Aguilera, Replication fork progression is impaired by transcription in hyperrecombinant yeast cells lacking a functional THO complex. *Mol. Cell Biol.* **26**, 3327-3334 (2006).
49. E. Grabczyk, M. Mancuso, M. C. Sammarco, A persistent RNA.DNA hybrid formed by transcription of the Friedreich ataxia triplet repeat in live bacteria, and by T7 RNAP in vitro. *Nucleic Acids Res.* **35**, 5351-5359 (2007).
50. Y. W. Lim, L. A. Sanz, X. Xu, S. R. Hartono, F. Chedin, Genome-wide DNA hypomethylation and RNA:DNA hybrid accumulation in Aicardi-Goutieres syndrome. *Elife* **4** (2015).
51. V. Bhatia, S. I. Barroso, M. L. Garcia-Rubio *et al.*, BRCA2 prevents R-loop accumulation and associates with TREX-2 mRNA export factor PCID2. *Nature* **511**, 362-365 (2014).
52. S. J. Hill, T. Rolland, G. Adelmant *et al.*, Systematic screening reveals a role for BRCA1 in the response to transcription-associated DNA damage. *Genes Dev* **28**, 1957-1975 (2014).
53. L. Wahba, J. D. Amon, D. Koshland, M. Vuica-Ross, RNase H and multiple RNA biogenesis factors cooperate to prevent RNA:DNA hybrids from generating genome instability. *Mol. Cell* **44**, 978-988 (2011).
54. X. Li, J. L. Manley, Inactivation of the SR protein splicing factor ASF/SF2 results in genomic instability. *Cell* **122**, 365-378 (2005).
55. P. Huertas, A. Aguilera, Cotranscriptionally formed DNA:RNA hybrids mediate transcription elongation impairment and transcription-associated recombination. *Mol. Cell* **12**, 711-721 (2003).

56. E. Y. Chang, C. A. Novoa, M. J. Aristizabal *et al.*, RECQ-like helicases Sgs1 and BLM regulate R-loop-associated genome instability. *J. Cell Biol.* **216**, 3991-4005 (2017).
57. A. Cristini, M. Groh, M. S. Kristiansen, N. Gromak, RNA/DNA Hybrid Interactome Identifies DXH9 as a Molecular Player in Transcriptional Termination and R-Loop-Associated DNA Damage. *Cell Rep.* **23**, 1891-1905 (2018).
58. D. Hodroj, B. Recolin, K. Serhal *et al.*, An ATR-dependent function for the Ddx19 RNA helicase in nuclear R-loop metabolism. *Embo. J.* **36**, 1182-1198 (2017).
59. S. C. Sridhara, S. Carvalho, A. R. Grosso *et al.*, Transcription Dynamics Prevent RNA-Mediated Genomic Instability through SRPK2-Dependent DDX23 Phosphorylation. *Cell Rep.* **18**, 334-343 (2017).
60. I. X. Wang, C. Grunseich, J. Fox *et al.*, Human proteins that interact with RNA/DNA hybrids. *Genome Res.* **28**, 1405-1414 (2018).
61. E. C. Friedberg, A history of the DNA repair and mutagenesis field The discovery of base excision repair. *DNA Repair* **37**, A35-A39 (2016).
62. A. P. Bertolin, S. F. Mansilla, V. Gottifredi, The identification of translesion DNA synthesis regulators: Inhibitors in the spotlight. *DNA Repair (Amst)* **32**, 158-164 (2015).
63. A. R. Lehmann, A. Niimi, T. Ogi *et al.*, Translesion synthesis: Y-family polymerases and the polymerase switch. *DNA Repair (Amst)* **6**, 891-899 (2007).
64. M. F. Goodman, Error-prone repair DNA polymerases in prokaryotes and eukaryotes. *Annu. Rev. Biochem.* **71**, 17-50 (2002).
65. L. S. Waters, B. K. Minesinger, M. E. Wiltrout *et al.*, Eukaryotic Translesion Polymerases and Their Roles and Regulation in DNA Damage Tolerance. *Microbiol. Mol. Biol. R* **73**, 134-+ (2009).
66. E. C. Friedberg, A. R. Lehmann, R. P. P. Fuchs, Trading places: How do DNA polymerases switch during translesion DNA synthesis? (vol 18, pg 499, 2005). *Mol. Cell* **19**, 143-143 (2005).
67. A. R. Lehmann, R. P. Fuchs, Gaps and forks in DNA replication: Rediscovering old models. *DNA Repair (Amst)* **5**, 1495-1498 (2006).
68. S. D. McCulloch, T. A. Kunkel, The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Res.* **18**, 148-161 (2008).
69. A. R. Lehmann, New functions for Y family polymerases. *Mol. Cell* **24**, 493-495 (2006).
70. P. L. Foster, G. Gudmundsson, J. M. Trimarchi, H. Cai, M. F. Goodman, Proofreading-defective DNA polymerase II increases adaptive mutation in Escherichia coli. *Proc Natl Acad Sci U S A* **92**, 7951-7955 (1995).
71. S. Rangarajan, G. Gudmundsson, Z. Qiu, P. L. Foster, M. F. Goodman, Escherichia coli DNA polymerase II catalyzes chromosomal and episomal DNA synthesis in vivo. *Proc Natl Acad Sci U S A* **94**, 946-951 (1997).
72. R. L. Frisch, Y. Su, P. C. Thornton *et al.*, Separate DNA Pol II- and Pol IV-dependent pathways of stress-induced mutation during double-strand-break repair in Escherichia coli are controlled by RpoS. *J. Bacteriol.* **192**, 4694-4700 (2010).
73. C. J. Kenyon, G. C. Walker, DNA-Damaging Agents Stimulate Gene-Expression at Specific Loci in Escherichia-Coli. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences* **77**, 2819-2823 (1980).
74. S. Mead, A. Vaisman, M. Valjavec-Gratian *et al.*, Characterization of polVR391: a Y-family polymerase encoded by rumA'B from the IncJ conjugative transposon, R391. *Mol. Microbiol.* **63**, 797-810 (2007).
75. M. F. Goodman, R. Woodgate, Translesion DNA polymerases. *Cold Spring Harb Perspect Biol* **5**, a010363 (2013).

76. R. E. Johnson, S. Prakash, L. Prakash, Efficient bypass of a thymine-thymine dimer by yeast DNA polymerase, Poleta. *Science* **283**, 1001-1004 (1999).
77. J. H. Yoon, L. Prakash, S. Prakash, Highly error-free role of DNA polymerase eta in the replicative bypass of UV-induced pyrimidine dimers in mouse and human cells. *Proc Natl Acad Sci U S A* **106**, 18219-18224 (2009).
78. J. E. Sale, A. R. Lehmann, R. Woodgate, Y-family DNA polymerases and their role in tolerance of cellular DNA damage. *Nat. Rev. Mol. Cell Bio.* **13**, 141-152 (2012).
79. B. F. Yuan, C. J. You, N. Andersen *et al.*, The Roles of DNA Polymerases kappa and iota in the Error-free Bypass of N-2-Carboxyalkyl-2 '-deoxyguanosine Lesions in Mammalian Cells. *J. Biol. Chem.* **286**, 17503-17511 (2011).
80. D. F. Jarosz, V. G. Godoy, J. C. Delaney, J. M. Essigmann, G. C. Walker, A single amino acid governs enhanced activity of DinB DNA polymerases on damaged templates. *Nature* **439**, 225-228 (2006).
81. J. A. Brown, J. D. Fowler, Z. C. Suo, Kinetic Basis of Nucleotide Selection Employed by a Protein Template-Dependent DNA Polymerase. *Biochemistry-us.* **49**, 5504-5510 (2010).
82. E. Mentegari, E. Crespan, L. Bavagnoli *et al.*, Ribonucleotide incorporation by human DNA polymerase eta impacts translesion synthesis and RNase H2 activity. *Nucleic Acids Res.* **45**, 2600-2614 (2017).
83. P. Sarkies, C. Reams, L. J. Simpson, J. E. Sale, Epigenetic Instability due to Defective Replication of Structured DNA. *Mol. Cell* **40**, 703-713 (2010).
84. C. M. Wickramasinghe, H. Arzouk, A. Frey, A. Maiter, J. E. Sale, Contributions of the specialised DNA polymerases to replication of structured DNA. *DNA Repair* **29**, 83-90 (2015).
85. E. Walsh, X. Wang, M. Y. Lee, K. A. Eckert, Mechanism of replicative DNA polymerase delta pausing and a potential role for DNA polymerase kappa in common fragile site replication. *J. Mol. Biol.* **425**, 232-243 (2013).
86. K. L. Burr, S. Velasco-Miguel, V. S. Duvvuri *et al.*, Elevated mutation rates in the germline of Polkappa mutant male mice. *DNA Repair (Amst)* **5**, 860-862 (2006).
87. R. Betous, L. Rey, G. L. Wang *et al.*, Role of TLS DNA Polymerases eta and kappa in Processing Naturally Occurring Structured DNA in Human Cells. *Mol. Carcinogen.* **48**, 369-378 (2009).
88. J. C. Delaney, J. M. Essigmann, Assays for determining lesion bypass efficiency and mutagenicity of site-specific DNA lesions in vivo. *DNA Repair, Pt A* **408**, 1-15 (2006).
89. B. F. Yuan, Y. S. Wang, Mutagenic and cytotoxic properties of 6-thioguanine, S-6-methylthioguanine, and guanine-S-6-sulfonic acid. *J. Biol. Chem.* **283**, 23665-23670 (2008).
90. C. J. You, Y. S. Wang, Mass Spectrometry-Based Quantitative Strategies for Assessing the Biological Consequences and Repair of DNA Adducts. *Accounts. Chem. Res.* **49**, 205-213 (2016).
91. C. J. You, Y. S. Wang, Quantitative measurement of transcriptional inhibition and mutagenesis induced by site-specifically incorporated DNA lesions in vitro and in vivo. *Nature Protocols* **10**, 1389-1406 (2015).
92. M. R. Wilkins, J. C. Sanchez, A. A. Gooley *et al.*, Progress with proteome projects: Why all proteins expressed by a genome should be identified and how to do it. *Biotechnology and Genetic Engineering Reviews, Vol 13* **13**, 19-50 (1996).
93. J. Parsons, C. Francavilla, 'Omics Approaches to Explore the Breast Cancer Landscape. *Frontiers in Cell and Developmental Biology* **7** (2020).

94. M. F. Tong, H. M. Liu, J. Y. Hao, D. M. Fan, Comparative pharmacoproteomics reveals potential targets for berberine, a promising therapy for colorectal cancer. *Biochem. Bioph. Res. Co* **525**, 244-250 (2020).
95. S. Saleem, S. Tariq, I. Aleem *et al.*, Proteomics analysis of colon cancer progression. *Clinical Proteomics* **16** (2019).
96. J. C. Rogers, R. D. Bomgarden, Sample Preparation for Mass Spectrometry-Based Proteomics; from Proteomes to Peptides. *Modern Proteomics - Sample Preparation, Analysis and Practical Applications* **919**, 43-62 (2016).
97. E. J. Dupree, M. Jayathirtha, H. Yorkey *et al.*, A Critical Review of Bottom-Up Proteomics: The Good, the Bad, and the Future of This Field. *Proteomes* **8** (2020).
98. J. D. Martell, T. J. Deerinck, Y. Sancak *et al.*, Engineered ascorbate peroxidase as a genetically encoded reporter for electron microscopy. *Nat. Biotechnol.* **30**, 1143-+ (2012).
99. H. W. Rhee, P. Zou, N. D. Udeshi *et al.*, Proteomic Mapping of Mitochondria in Living Cells via Spatially Restricted Enzymatic Tagging. *Science* **339**, 1328-1331 (2013).
100. A. L. Richards, M. Eckhardt, N. J. Krogan, Mass spectrometry-based protein-protein interaction networks for the study of human diseases. *Molecular Systems Biology* **17** (2021).
101. I. F. Chang, Mass spectrometry-based proteomic analysis of the epitope-tag affinity purified protein complexes in eukaryotes. *Proteomics* **6**, 6158-6166 (2006).
102. E. L. Huttlin, L. Ting, R. J. Bruckner *et al.*, The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162**, 425-440 (2015).
103. E. L. Huttlin, R. J. Bruckner, J. A. Paulo *et al.*, Architecture of the human interactome defines protein communities and disease networks. *Nature* **545**, 505-+ (2017).
104. M. Dalvai, J. Loehr, K. Jacquet *et al.*, A Scalable Genome-Editing-Based Approach for Mapping Multiprotein Complexes in Human Cells. *Cell Rep.* **13**, 621-633 (2015).
105. A. L. Richards, M. Eckhardt, N. J. Krogan, Mass spectrometry-based protein-protein interaction networks for the study of human diseases. *Mol Syst Biol* **17**, e8792 (2021).
106. L. M. Smith, J. Z. Sanders, R. J. Kaiser *et al.*, Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674-679 (1986).
107. Z. G. Chidgeavadze, R. S. Beabealashvilli, A. M. Atrazhev *et al.*, 2',3'-Dideoxy-3' aminonucleoside 5'-triphosphates are the terminators of DNA synthesis catalyzed by DNA polymerases. *Nucleic Acids Res.* **12**, 1671-1686 (1984).
108. J. Shendure, H. Ji, Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135-1145 (2008).
109. P. J. Park, ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **10**, 669-680 (2009).
110. P. Wulfridge, K. Sarma, A nuclease- and bisulfite-based strategy captures strand-specific R-loops genome-wide. *Elife* **10** (2021).
111. J. Nadel, R. Athanasiadou, C. Lemetre *et al.*, RNA:DNA hybrids in the human genome have distinctive nucleotide characteristics, chromatin composition, and transcriptional relationships. *Epigenetics & Chromatin* **8** (2015).
112. L. A. Sanz, S. R. Hartono, Y. W. Lim *et al.*, Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals. *Mol. Cell* **63**, 167-178 (2016).
113. K. Wang, H. H. Wang, C. H. Li *et al.*, Genomic profiling of native R loops with a DNA-RNA hybrid recognition sensor. *Science Advances* **7** (2021).
114. J. G. Dumelie, S. R. Jaffrey, Defining the location of promoter associated R-loops at near-nucleotide resolution using bisDRIP-seq. *Elife* **6** (2017).

115. T. Wu, R. T. Lyu, Q. C. You, C. He, Kethoxal-assisted single-stranded DNA sequencing captures global transcription dynamics and enhancer activity in situ (vol 17, pg 515, 2020). *Nature Methods* **17**, 749-749 (2020).
116. X. C. Weng, J. Gong, Y. Chen *et al.*, Keth-seq for transcriptome-wide RNA structure mapping. *Nat. Chem. Biol.* **16**, 489-+ (2020).
117. T. Wu, R. Lyu, C. He, spKAS-seq reveals R-loop dynamics using low-input materials by detecting single-stranded DNA with strand specificity. *Science Advances* **8** (2022).

Chapter 2

DNA Polymerase II Supports the Replicative Bypass of N2-Alkyl-2′-deoxyguanosine Lesions in Escherichia coli Cells

Introduction

Cells are continuously exposed to alkylating agents that can damage DNA, and the N^2 position of guanine is a common site of alkylation (1,2). For instance, the N^2 of guanine is susceptible to reaction with formaldehyde and acetaldehyde, which can be produced endogenously or present in external sources, including diesel exhaust, cigarette smoke, *etc* (3,4). In addition, methylglyoxal, an air pollutant and a byproduct of endogenous glycolysis, can induce the stable N^2-(1-carboxyethyl)-2′-deoxyguanosine (N^2-CE-dG) in DNA (5).

DNA adducts, if left unrepaired, can block DNA replication and induce nucleobase substitutions in that process; if mutations occur in oncogenes or tumor suppressor genes, these adducts may contribute to carcinogenesis (5-7). To ameliorate the genotoxic effects of DNA lesions, cells are equipped with various DNA repair machineries to remove these lesions. Cells are also evolved with DNA damage tolerance pathways to cope with unrepaired lesions, where translesion synthesis (TLS) is one of these pathways for overcoming replication blockage conferred by DNA lesions (1,8,9). In *Escherichia coli*, three TLS polymerases (Pol II, Pol IV, and Pol V) can be induced under SOS conditions; Pol II and Pol IV participate in mainly error-free TLS of specific DNA lesions, whereas Pol V can bypass a wide range of DNA lesions in a more error-prone manner (10).

Several studies have been conducted to examine the effects of N^2-alkyl-dG adducts on DNA replication in cells. Yuan *et al.* (11) showed that the *R* diastereomer of N^2-CE-dG blocks DNA replication more strongly than the *S* diastereomer in *E. coli*, and replicative bypass of these

lesions is largely error-free, where Pol IV is required for the accurate and efficient bypass of these lesions. In addition, faithful replication of N^2-CE-dG and a number of simple N^2-alkyl-dG lesions (alkyl = Me, Et, *n*Pr, and *n*Bu) in mammalian cells requires polymerases κ and ι, where Pol κ is the mammalian ortholog of *E. coli* Pol IV (11,12). On the other hand, Shrivastav et al. (13) found that the replication across N^2-Me-dG and N^2-Et-dG is accurate and efficient in *E. coli* cells; depletion of Pol IV, however, does not perturb the efficiency or fidelity of replication across these lesions. It has not yet been investigated systematically how simple N^2-alkyl-dG lesions influence DNA replication in *E. coli* cells or how the three SOS-induced DNA polymerases modulate the replication across these lesions. To answer these questions, we examined the efficiencies and fidelities of replication across four N^2-alkyl-dG lesions with different sizes and various structures of alkyl groups (N^2-Et-dG, N^2-*n*Bu-dG, N^2-*i*Bu-dG, and N^2-*s*Bu-dG, Figure 2-7) in *E. coli* cells and the functions of the SOS-induced DNA polymerases in supporting their replication bypass.

We utilized a modified competitive replication and adduct bypass (CRAB) assay to explore how replication efficiency and fidelity of a single-stranded M13 plasmid are influenced by the presence of site-specifically inserted N^2-alkyl-dG lesions (Figures 2-1–2-5 and Figure 2-8) (11). In brief, lesion-containing single-stranded M13 plasmids were mixed individually with a lesion-free competitor plasmid at a specific molar ratio and allowed to replicate in SOS-induced AB1157 *E. coli* cells that are proficient in TLS or with one or more SOS-induced DNA polymerases being genetically depleted. In this respect, the competitor genome, which harbors three more nucleotides but is lesion-free, acts as an internal standard for measuring the replication efficiency across the damage site. After progeny genome isolation, PCR amplification, and restriction digestion, the released oligodeoxyribonucleotides (ODNs) were analyzed by LC-MS/MS and native PAGE to identify and quantify the replication products.

Methods and Materials

All chemicals, unless otherwise specified, were from Sigma-Aldrich (St. Louis, MO), and all enzymes were from New England Biolabs (Ipswich, MA). 1,1,1,3,3,3-Hexafluoro-2-propanol, [γ-^{32}P]ATP, and unmodified oligodeoxyribonucleotides (ODNs) were obtained from Oakwood Products Inc. (West Columbia, SC), PerkinElmer Life Sciences (Piscataway, NJ), and Integrated DNA Technologies (Coralville, IA) respectively. The 12-mer N^2-alkyl-dG-containing ODNs employed in this study were synthesized following previously published procedures, where the 12-mer N^2-Et-dG- and N^2-*n*Bu-dG-harboring ODNs were characterized previously.

M13mp7(L2) plasmid, wild-type AB1157 E. coli strains were kindly provided by Prof. John M. Essigmann. Polymerase-deficient AB1157 strains [Δpol B1::spec (Pol II deficient), ΔdinB (Pol IV-deficient), ΔumuC::kan (Pol V deficient) and ΔdinB ΔumuC::kan (Pol IV, Pol V-double knockout) Δpol B1::spec ΔdinB ΔumuC::kan (Pol II, Pol IV, Pol V-triple knockout)] were generously provided by Prof. Graham C. Walker .

Construction of lesion-containing and lesion-free M13 genomes

The lesion-containing and lesion-free control/competitor genomes were prepared following previously published procedures. Briefly, the 12-mer N^2-alkyl-dG-containing ODNs were 5′- phosphorylated and ligated with a 10-mer lesion-free ODN (5′-ACTGGAAGAC-3′) and the resulting 22-mer ODNs were purified by denaturing polyacrylamide gel electrophoresis (PAGE). M13 plasmid (20 pmol) was treated with EcoRI-HF at room temperature overnight, and the ensuing linearized vector was mixed with 2 scaffolds (25 pmol each), each spanning one end of the linear vector. To the mixture was subsequently added 30 pmol of 5′-phosphorylated 22-mer lesion-containing ODN or 25-mer competitor ODN (5′-GCAGGATGTCATGGCGATAAGCTAT-3′). The resulting mixture was treated with T4 DNA ligase at 16°C for 8 h, followed by

incubation with T4 DNA polymerase at 16°C for 1 h to remove the residual unligated plasmid and excess scaffold ODNs. The lesion-containing and lesion-free plasmids were purified using Cycle Pure Kit (Omega) and subsequently normalized against the competitor plasmid by denaturing PAGE.

Transformation of mixed plasmids into E. coli cells

The lesion-free and lesion-containing genomes were mixed with the competitor genome at molar ratios of 1:1 and 3:1, respectively. The mixtures were transfected into SOS-induced, electrocompetent AB1157 *E*. coli strain as well as the isogenic cells with Pol II, Pol IV, and Pol V being individually knocked out, or with Pol IV and Pol V or all three polymerases being simultaneously depleted. The SOS induction was executed by exposing *E*. coli cells with 254 nm ultraviolet light at a dose of 45 J/m^2. The *E*. coli cells were subsequently cultured in lysogeny broth (LB) medium at 37°C for 5.5 h. The phage was isolated from the supernatant by centrifugation, and then transfected into SCS110 cells to increase further the progeny/lesion-genome ratio. The resulting phage was purified by Qiaprep Spin M13 Kit (Qiagen) to obtain the ssM13 progeny for PCR amplification.

PCR amplification and PAGE quantification

A modified version of the competitive replication and adduct bypass (CRAB) assay was employed to assess how N^2-alkyl-dG lesions affect the efficiencies and fidelities of DNA replication in *E*. coli cells. After cellular replication, the single-stranded M13 progeny genome was amplified by using previously reported PCR conditions. For determining the bypass efficiencies, 50 ng of PCR products were mixed with 10 U BbsI-HF, 10 U recombinant shrimp alkaline phosphatase (rSAP) and CutSmart buffer (New England Biolabs) to give a total volume of 10 μl and incubated at 37°C for 25 min, followed by heating at 80°C for 10 min to deactivate

the enzymes. To the above digested PCR products were added a 5-μl mixture of 5 mM DTT, 1.66 pmol [γ-^{32}P]ATP, 10 U T4 polynucleotide kinase (T4 PNK), T4 PNK buffer and water, and the mixture was incubated at 37°C for 30 min. T4 PNK was then deactivated by heating at 80°C for 10 min. To the resulting solution were added a 5-μl mixture of 10 U MluCI, Cutsmart buffer and water, and the mixture was incubated at 37°C for 25 min. The reaction was finally quenched by adding a 15-μl 2× formamide gel loading buffer. The above restriction digestion and post-labeling procedures yielded a 13-mer 5′- ^{32}P-labeled fragment for the progeny of the competitor plasmid and 10-mer 5′- ^{32}P-labeled fragment for the progenies of the control or lesion-carrying plasmids that are free of deletion and insertion mutations. The mixture was resolved by using 30% native polyacrylamide gel (acrylamide:bis-acrylamide = 19:1), and the gel band intensities were measured by phosphorimaging analysis. The bypass efficiency was calculated by using the following equation: Bypass efficiency (%) = (lesion signal/competitor signal)/(control signal/competitor signal) × 100%.

Identification of mutagenic products by LC–MS and MS/MS

Approximately 3.0 μg of PCR products were incubated with 50 U BbsI-HF restriction endonuclease and 20 U rSAP in 250 μl CutSmart buffer at 37°C for 2 h, and subsequently deactivated by heating at 80°C for 20 min. To the mixture was added 50 U MluCI and incubated at 37°C for another 1 h. The reaction solution was then extracted with phenol/chloroform/isoamyl alcohol (25:24:1, v/v) and the top aqueous layer was collected and evaporated to dryness. The residues were then dissolved in 100 μl water and desalted by using Waters Oasis HLB extraction cartridge (Milford, MA). A 10-μl aliquot was injected for LC–MS/MS analysis on an LTQ linear ion trap mass spectrometer (Thermo Electron, San Jose, CA, USA). An Agilent Zorbax SB-C18 column (0.5 × 250 mm, 5 μm in particle size) was employed. The mass spectrometer was set up

for monitoring the fragmentation of the $[M–3H]^{3–}$ ions of 10-mer d(GGCGXGCTAT), with 'X' being 'A', 'T', 'C' or 'G'. The fragment ions detected in MS/MS were manually assigned.

Results

The results from native PAGE analyses of the ensuing radiolabeled fragments revealed that no mutagenic products were induced by any of the N^2-alkyl-dG lesions, and a similar finding was made from LC-MS/MS analysis of the corresponding non-radiolabeled restriction fragments (Figures 2-3–2-5 and Figure 2-8B, C). We also determined the bypass efficiencies of the N^2-alkyl-dG lesions by measuring the ratio of intensity of the 10 mer band arising from the lesion-containing genome over that of the 13 mer band emanating from the lesion-free competitor genome and further normalizing the ratio to that observed for the control dG-containing genome. The results showed that all four N^2-alkyl-dG lesions could impede DNA replication in SOS-induced *E. coli* cells, with the bypass efficiencies for N^2-Et-dG, N^2-nBu-dG, N^2-*i*Bu-dG, and N^2-*s*Bu-dG being 36.5, 31.5, 27.1, and 28.5%, respectively. Hence, the three N^2-Bu-dG lesions display blockage effects on DNA replication slightly stronger than that with N^2-Et-dG; the replication efficiencies across the three N^2-Bu-dG lesions are, nonetheless, not appreciably affected by the structures of the butyl groups (*n*Bu, *i*Bu, and *s*Bu) (Figure 2-8D).

It is worth noting that the bypass efficiency for N^2-Et-dG observed in the present study is lower than what was observed previously by Shrivastav *et al.* (13), which could be attributed to different flanking sequences of the lesion employed in the two studies. In this vein, sequence contexts surrounding DNA damage sites are known to modulate the efficiencies and fidelities of DNA replication across these sites (14, 15). The bypass efficiencies for N^2-Et-dG and N^2-*n*Bu-dG obtained from this study were also lower than what Wu et al. reported recently for the same

lesions in HEK293T human embryonic kidney cells, which could be due to differences in replication machineries in E. *coli* and human cells (12).

We next examined the roles of Pol II, Pol IV, and Pol V in bypassing the N^2-alkyl-dG adducts by performing the replication experiments using the SOS-induced isogenic *E. coli* strains where these polymerases were individually or concurrently ablated. Except that depletion of Pol IV alone did not give rise to significant drops in bypass efficiencies for N^2-Et-dG or N^2-*n*Bu-dG, individual depletion of each of the three SOS-induced DNA polymerases led to substantial attenuations in bypass efficiencies for all four N^2-alkyl-dG lesions, with the most pronounced decreases being observed for the Pol II-deficient background (Figure 2-8D). Additional drops in bypass efficiencies were observed for the three N^2-Bu-dG lesions in Pol IV and Pol V double knockout background compared to those with depletion of either polymerase alone (Figure 2-8D).

For comparison, we also examined how these lesions modulate the replicative bypass of these lesions in wild-type and Pol II-depleted AB1157 cells without SOS induction. Our results showed that SOS induction led to augmented bypass efficiencies for all four lesions in AB1157 cells (Figure 2-8D, E). Depletion of Pol II, however, does not alter appreciably the bypass efficiencies for any of the four N^2-alkyl-dG lesions in AB1157 cells without SOS induction (Figure 2-8E). It was shown previously that SOS induction could give rise to a 7-fold elevation in expression level of Pol II (16). Thus, the lack of effect of Pol II deletion on the bypass efficiencies of these lesions in uninduced *E. coli* cells could be due to the relatively low level of expression of Pol II in wild-type AB1157 cells without SOS induction.

Discussion

Exposure to endogenous and exogenous genotoxic agents can lead to the formation of various DNA lesions, many of which block replicative DNA polymerases and require TLS

polymerases for their replicative bypass (17). A large body of literature revealed the roles of B- and Y-family polymerases in modulating the cytotoxic and mutagenic properties of various DNA lesions (18, 19). In this vein, Pol IV and Pol V were found to participate in error-free TLS and induce a -1 frameshift mutation at the site of an N^2-dG adduct of benzo[*a*]pyrene; Pol II, however, is involved in bypassing the bulky N^2-dG adduct of acetylaminofluorene and elicits a -2 frameshift mutation at the lesion site (20, 21). In addition, previous in vitro biochemical studies showed that purified Pol IV preferentially inserts a dCMP opposite N^2-Et-dG, N^2-*i*Bu-dG, and N^2-CE-dG in template DNA (11, 22), though the kinetic parameters for nucleotide insertions opposite simple N^2-alkyl-dG lesions were not measured. It will be important to determine, in the future, the efficiencies and fidelities of Pol II- and Pol IV-catalyzed nucleotide incorporation opposite the N^2-alkyl-dG lesions.

The major finding from the present study is about the contributions of the three SOS-induced DNA polymerases, namely, Pol II, Pol IV, and Pol V, in bypassing four minor-groove N^2-alkyl-dG adducts (alkyl = Et, *n*Bu, *i*Bu, and *s*Bu) in *E.* coli cells. Similar to what were observed for N^2-furfuryl-dG and N^2-tetrahydrofuran-2-yl-methyl-dG lesions (13), we found that the smaller N^2-Et-dG and N^2-Bu-dGs were not mutagenic in AB1157 cells or any of the isogenic polymerase-deficient strains tested. Our results support that all three TLS polymerases in *E. coli* are involved in the error-free TLS of these four lesions. A small alkyl group adducted to the N^2 of guanine does not strongly impair the base-pairing property of the nucleobase, whereas those adducts with large alkyl groups can induce DNA frameshifts or single-base substitutions during DNA replication (20). Meanwhile the loss of Pol II results in the most marked diminutions in bypass efficiencies in SOS-induced AB1157 cells, underscoring the major role of this polymerase in overcoming the replication blockage imposed by these N^2-alkyl-dG lesions.

It is worth noting that Pol IV was found to be the major DNA polymerase responsible for bypassing N^2-CE-dG lesions *in vivo* (11), whereas Pol II was the main polymerase involved in bypassing the N^2-dG lesions with small alkyl groups. The exact reason for these differences remains unclear and awaits further investigation. In this vein, the X-ray crystal structure of a ternary complex of *E. coli* Pol II, duplex DNA, and an incoming dCTP showed that the active site of the polymerase facing the minor-groove N^2 position of template dG is quite spacious (23), which should be able to accommodate alkyl groups adducted to the N^2 of dG (Figure 2-6). This may explain Pol II's role in supporting the accurate and efficient bypass of N^2-alkyl-dG lesions. In this context, it is of note that minor-groove O^2-alkyl-dT lesions are highly mutagenic in *E. coli* cells and their efficient bypass mainly requires Pol V (24).

These previously published results, together with the observations made from the present study, indicate that all three TLS polymerases can participate in bypassing minor-groove lesions; Pol V tends to transverse, in an error-prone manner, those lesions with the hydrogen bonding properties of the nucleobases being disrupted (e.g. the O^2-alkyl-dT lesions), whereas Pol II and Pol IV tend to bypass accurately minor-groove N^2-alkyl-dG lesions. Moreover, DNA Pol II, which is B-family DNA polymerase, was also found to participate in the error-free bypass of a major-groove N^6-benz[*a*]anthracene adenine adduct (20). Along this line, several B-family polymerases were shown to possess a conserved motif that scans the DNA minor groove for lesions and misincorporations (25). Therefore, Pol II is capable of bypassing accurately both major- and minor-groove alkylated DNA lesions. Together, the results from previous studies and the current work reveal that nuances of TLS can be modulated by the subtle differences in chemical structures of DNA lesions.

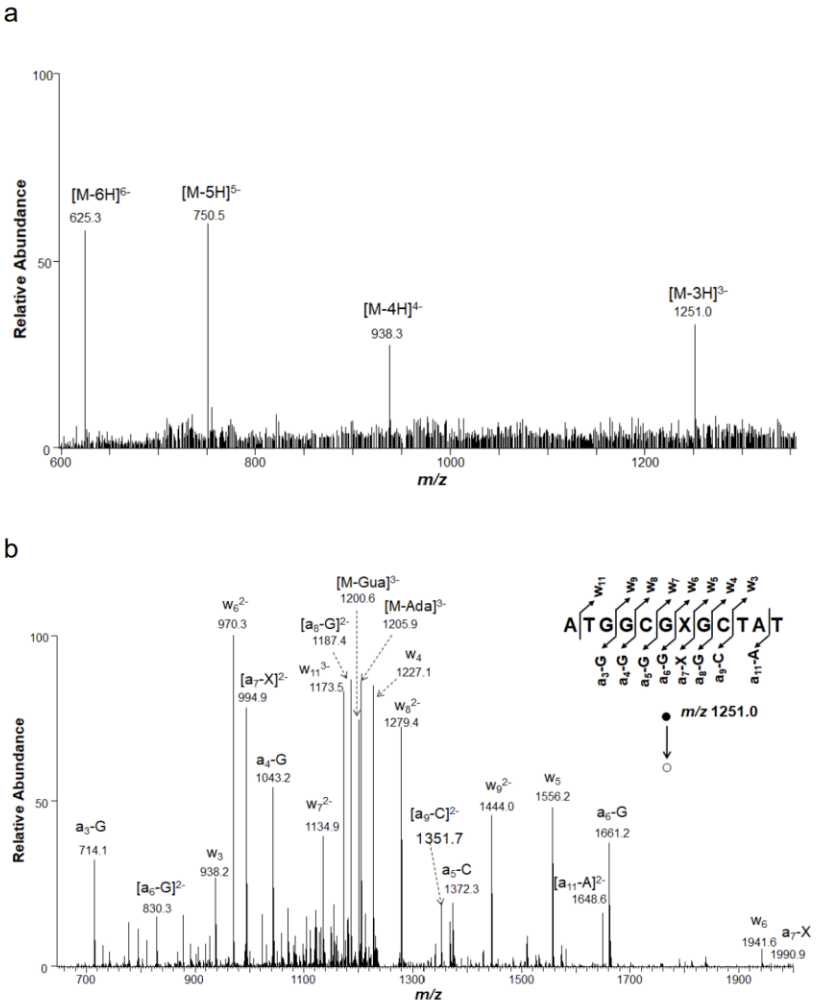

Figure 2-1. ESI-MS & MS/MS characterizations of d(ATGGCGXGCTAT), X= N^2-*i*Bu-dG. (a)Negative-ion ESI-MS; (b) product-ion spectrum of the $[M-3H]^{3-}$ ion (m/z 1251.0).

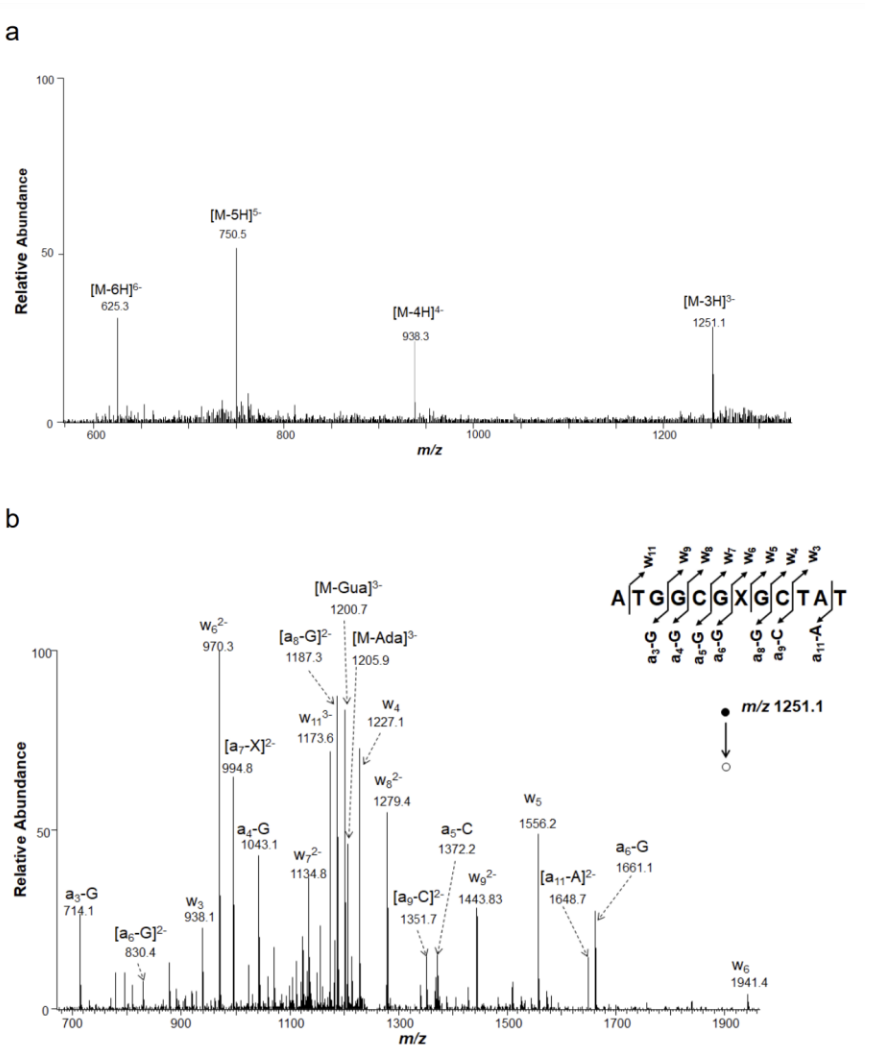

Figure 2-2. ESI-MS & MS/MS characterizations of d(ATGGCGXGCTAT), X= N^2-sBu-dG. (a) Negative-ion ESI-MS; (b) product-ion spectrum of the $[M-3H]^{3-}$ ion (m/z 1251.1).

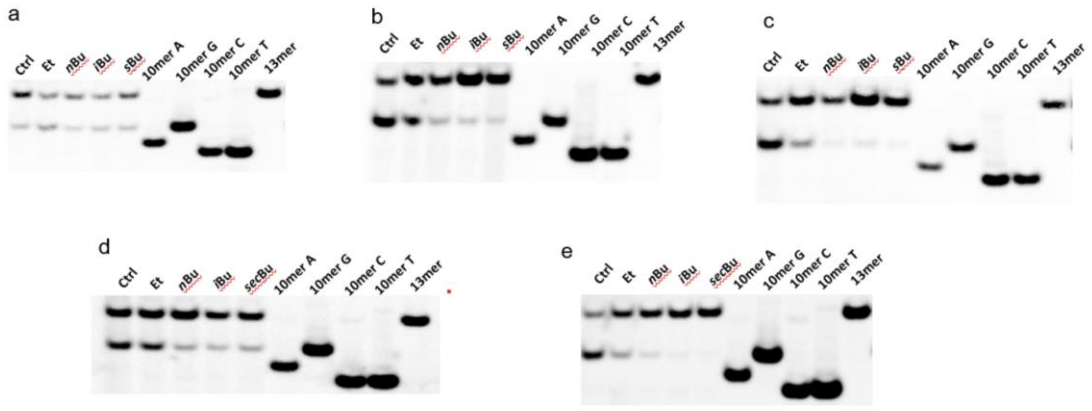

Figure 2-3. Native PAGE (30%) for monitoring the bypass efficiencies and mutation frequencies of N^2-alkyl-dG lesions in AB1157 cells or the isogenic cells with Pol II, Pol IV and Pol V being individually or simultaneously depleted. (A) AB1157, (B) SOS–Δ Pol IV, (C) SOS-ΔPol V, (D) SOS -DKO of Pol IV and Pol V, (E) SOS-TKO.

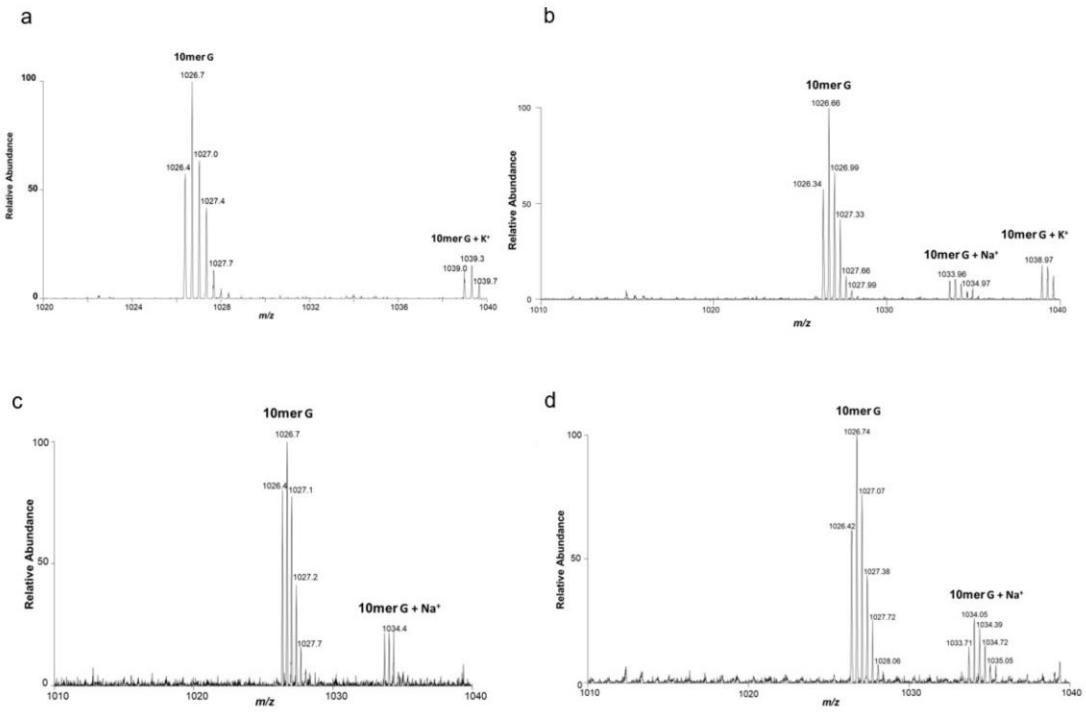

Figure 2-4. Higher-resolution "ultra-zoom scan" ESI-MS of the restriction fragments for the PCR products from the replication of single-stranded M13 genomes harboring a site-specifically inserted (A) N^2-EtdG, (B) N^2-*n*Bu*dG*, (C) N^2-*i*BudG, and (D) N^2-*s*BudG in AB1157 cells. Displayed in this figure are the $[M-3H]^{3-}$ ions of the strand initially contained the lesion.

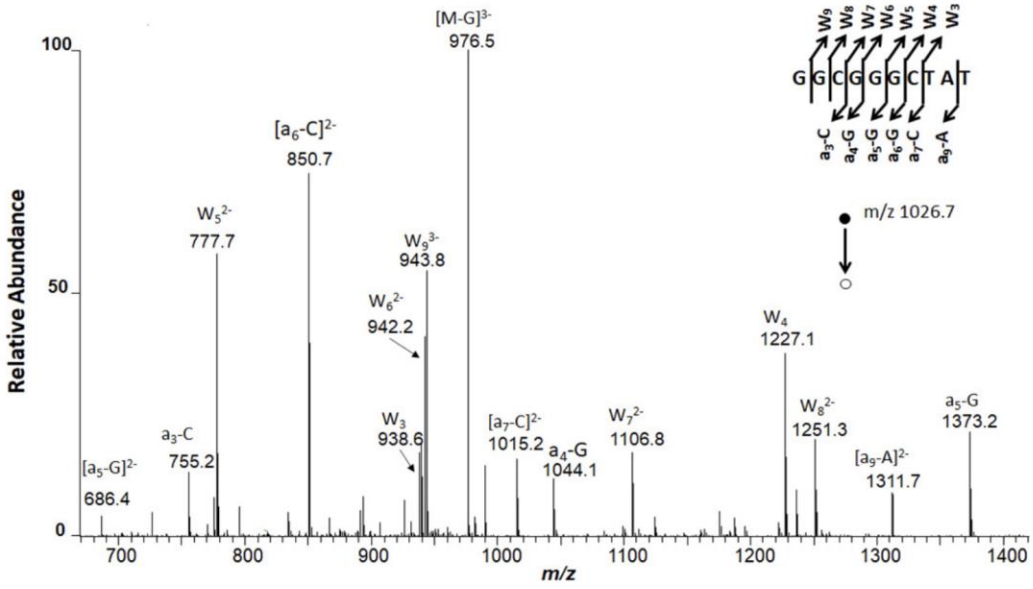

Figure 2-5. MS/MS for the identification of restriction fragments of PCR products. MS/MS for the $[M-3H]^{3-}$ ions of 10mer 5′-GGCGGGCTAT-3′ replication product of N^2-*n*Bu-dG–containing plasmid in Pol II deficient cells.

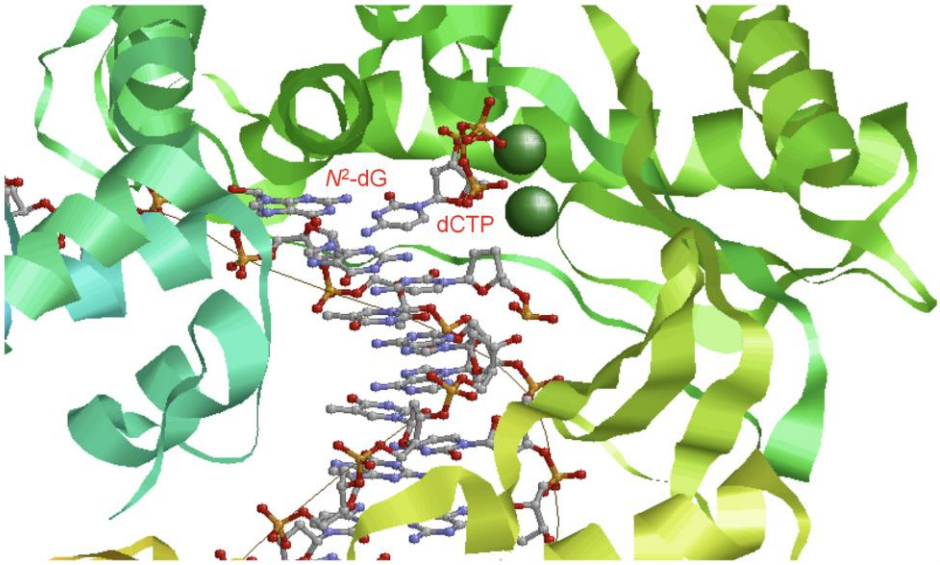

Figure 2-6. Partial structure of a ternary complex of E. coli DNA polymerase II, unmodified duplex DNA with a dG at the active site, and an incoming dCTP (3K59.pdb). The protein is shown in ribbon model, DNA and incoming dCTP are depicted in ball and stick model, and the Mg^{2+} ions are displayed in spacefill model.

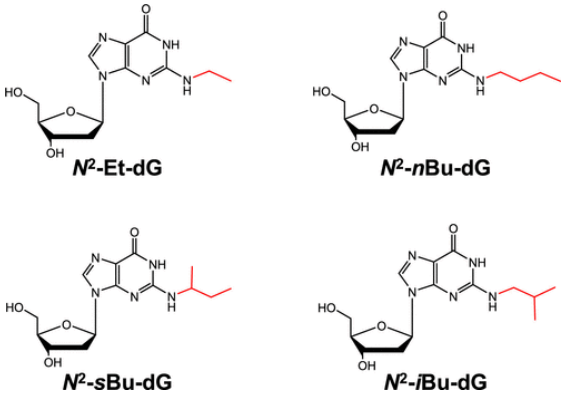

Figure 2-7. N^2-Alkyl-dG lesions investigated in this study.

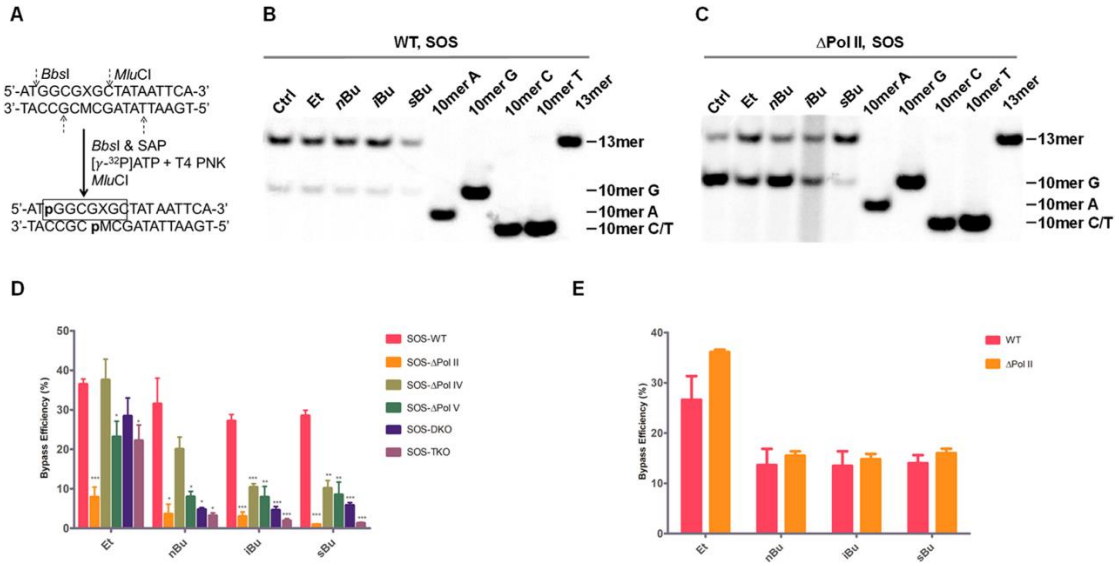

Figure 2-8. Restriction digestion and post labeling method for determining the cytotoxic and mutagenic properties of N^2-alkyl-dG lesions in wild-type AB1157 *E. coli* and the isogenic strains deficient in one or more SOS-induced DNA polymerases. (A) Restriction digestion and selective radiolabeling of the original lesion-containing strand or its complementary strand. (B,C) Representative gel images showing the *Bbs*I/*Mluc*I-produced restriction fragments of interest from the PCR products of progeny genomes of the indicated lesion- or control dG-containing plasmids isolated from SOS-induced wild-type (WT) and Pol II-deficient AB1157 *E. coli* cells. (D) Bypass efficiencies of N^2-alkyl-dG lesions in SOS-induced wild-type AB1157 cells and isogenic polymerase-deficient cells. (E) Bypass efficiencies of N^2-alkyl-dG lesions in wild-type and Pol II-deficient AB1157 E. coli cells without SOS induction. The bypass efficiency was calculated by using the following equation: bypass efficiency (%) = (10 mer lesion signal/13 mer competitor signal)/(10 mer control signal/13 mer competitor signal) × 100%. The data represents the mean ± SEM of results acquired from three independent replication experiments. *, $0.01 \leq P < 0.05$; **, $0.001 \leq P < 0.01$; ***, $P < 0.001$. The P values in (D) were calculated using a two-tailed, unpaired t test, and the values refer to the comparisons between SOS-induced WT and the isogenic TLS polymerase-deficient cells (listed above the columns for the polymerase-deficient cells).

References

1. L. R. Ferguson, H. Chen, A. R. Collins *et al.*, Genomic instability in human cancer: Molecular insights and opportunities for therapeutic attack and prevention through diet and nutrition. *Semin. Cancer Biol.* **35**, S5-S24 (2015).
2. T. Abbas, M. A. Keaton, A. Dutta, Genomic instability in cancer. *Cold Spring Harb Perspect Biol* **5**, a012914 (2013).
3. K. W. Caldecott, M. E. Ward, A. Nussenzweig, The threat of programmed DNA damage to neuronal genome integrity and plasticity. *Nat Genet* **54**, 115-120 (2022).
4. R. Madabhushi, L. Pan, L. H. Tsai, DNA damage and its links to neurodegeneration. *Neuron* **83**, 266-282 (2014).
5. C. E. Cunningham, M. J. MacAuley, G. Yadav *et al.*, Targeting the CINful genome: Strategies to overcome tumor heterogeneity. *Prog Biophys Mol Biol* **147**, 77-91 (2019).
6. S. Negrini, V. G. Gorgoulis, T. D. Halazonetis, Genomic instability--an evolving hallmark of cancer. *Nat Rev Mol Cell Biol* **11**, 220-228 (2010).
7. L. J. Marnett, J. P. Plastaras, Endogenous DNA damage and mutation. *Trends Genet* **17**, 214-221 (2001).
8. R. De Bont, N. van Larebeke, Endogenous DNA damage in humans: a review of quantitative data. *Mutagenesis* **19**, 169-185 (2004).
9. M. Katerji, P. J. Duerksen-Hughes, DNA damage in cancer development: special implications in viral oncogenesis. *Am J Cancer Res* **11**, 3956-3979 (2021).
10. K. Lu, S. Craft, J. Nakamura, B. C. Moeller, J. A. Swenberg, Use of LC-MS/MS and stable isotopes to differentiate hydroxymethyl and methyl DNA adducts from formaldehyde and nitrosodimethylamine. *Chem. Res. Toxicol.* **25**, 664-675 (2012).
11. J. D. Mcghee, P. H. V. Hippel, Formaldehyde as a Probe of DNA Structure .2. Reaction with Endocyclic Imino Groups of DNA Bases. *Biochemistry-us.* **14**, 1297-1303 (1975).
12. N. Shrivastav, B. I. Fedeles, D. Y. Li *et al.*, A Chemical Genetics Analysis of the Roles of Bypass Polymerase DinB and DNA Repair Protein AlkB in Processing N-2-Alkylguanine Lesions In Vivo. *Plos One* **9** (2014).
13. P. J. Brooks, J. A. Theruvathu, DNA adducts from acetaldehyde: implications for alcohol-related carcinogenesis. *Alcohol.* **35**, 187-193 (2005).
14. M. Y. Wang, N. X. Yu, L. Chen *et al.*, Identification of an acetaldehyde adduct in human liver DNA and quantitation as N-2-ethyldeoxyguanosine. *Chem. Res. Toxicol.* **19**, 319-324 (2006).
15. P. J. Thornalley, The Glyoxalase System - New Developments Towards Functional-Characterization of a Metabolic Pathway Fundamental to Biological Life. *Biochem. J.* **269**, 1-11 (1990).
16. L. A. Fothergill-Gilmore, P. A. Michels, Evolution of glycolysis. *Prog Biophys Mol Biol* **59**, 105-235 (1993).
17. P. J. Thornalley, Pharmacology of methylglyoxal: Formation, modification of proteins and nucleic acids, and enzymatic detoxification - A role in pathogenesis and antiproliferative chemotherapy. *Gen. Pharmacol-vasc. S* **27**, 565-573 (1996).
18. M. Frischmann, C. Bidmon, J. Angerer, M. Pischetsrieder, Identification of DNA adducts of methylglyoxal. *Chem. Res. Toxicol.* **18**, 1586-1592 (2005).
19. D. H. Phillips, Fifty years of benzo(a)pyrene. *Nature* **303**, 468-472 (1983).
20. S. S. Hecht, Tobacco smoke carcinogens and lung cancer. *J Natl Cancer Inst* **91**, 1194-1210 (1999).

21. C. E. Bostrom, P. Gerde, A. Hanberg *et al.*, Cancer risk assessment, indicators, and guidelines for polycyclic aromatic hydrocarbons in the ambient air. *Environ Health Perspect* **110 Suppl 3**, 451-488 (2002).
22. B. Bukowska, P. Sicinska, Influence of Benzo(a)pyrene on Different Epigenetic Processes. *Int J Mol Sci* **22** (2021).
23. D. Li, P. F. Firozi, L. E. Wang *et al.*, Sensitivity to DNA damage induced by benzo(a)pyrene diol epoxide and risk of lung cancer: a case-control analysis. *Cancer Res.* **61**, 1445-1450 (2001).
24. E. Kriek, M. Rojas, K. Alexandrov, H. Bartsch, Polycyclic aromatic hydrocarbon-DNA adducts in humans: relevance as biomarkers for exposure and cancer risk. *Mutat Res* **400**, 215-231 (1998).
25. E. Gyorffy, L. Anna, Z. Gyori *et al.*, DNA adducts in tumour, normal peripheral lung and bronchus, and peripheral blood lymphocytes from smoking and non-smoking lung cancer patients: correlations between tissues and detection by 32P-postlabelling and immunoassay. *Carcinogenesis* **25**, 1201-1209 (2004).
26. B. Bukowska, K. Mokra, J. Michalowicz, Benzo[a]pyrene-Environmental Occurrence, Human Exposure, and Mechanisms of Toxicity. *Int J Mol Sci* **23** (2022).
27. A. L. Piberger, C. T. Kruger, B. M. Strauch, B. Schneider, A. Hartwig, BPDE-induced genotoxicity: relationship between DNA adducts, mutagenicity in the in vitro PIG-A assay, and the transcriptional response to DNA damage in TK6 cells. *Arch. Toxicol.* **92**, 541-551 (2018).
28. B. Ma, P. W. Villalta, A. T. Zarth *et al.*, Comprehensive High-Resolution Mass Spectrometric Analysis of DNA Phosphate Adducts Formed by the Tobacco-Specific Lung Carcinogen 4-(Methylnitrosamino)-1-(3-pyridyl)-1-butanone. *Chem. Res. Toxicol.* **28**, 2151-2159 (2015).
29. D. Fu, J. A. Calvo, L. D. Samson, Balancing repair and tolerance of DNA damage caused by alkylating agents. *Nat Rev Cancer* **12**, 104-120 (2012).
30. F. Drablos, E. Feyzi, P. A. Aas *et al.*, Alkylation damage in DNA and RNA--repair mechanisms and medical significance. *DNA Repair (Amst)* **3**, 1389-1407 (2004).
31. Y. Tan, S. Guo, J. Wu *et al.*, DNA Polymerase eta Promotes the Transcriptional Bypass of N(2)-Alkyl-2'-deoxyguanosine Adducts in Human Cells. *J Am Chem Soc* **143**, 16197-16205 (2021).
32. J. Wu, H. Du, L. Li *et al.*, The Impact of Minor-Groove N(2)-Alkyl-2'-deoxyguanosine Lesions on DNA Replication in Human Cells. *ACS Chem. Biol.* **14**, 1708-1716 (2019).
33. Y. A. Hegazy, C. M. Fernando, E. J. Tran, The balancing act of R-loop biology: The good, the bad, and the ugly. *J. Biol. Chem.* **295**, 905-913 (2020).
34. I. M. Okazaki, K. Kinoshita, M. Muramatsu, K. Yoshikawa, T. Honjo, The AID enzyme induces class switch recombination in fibroblasts. *Nature* **416**, 340-345 (2002).
35. C. C. So, A. Martin, DSB structure impacts DNA recombination leading to class switching and chromosomal translocations in human B cells. *PLoS Genet* **15**, e1008101 (2019).
36. P. A. Ginno, P. L. Lott, H. C. Christensen, I. Korf, F. Chedin, R-Loop Formation Is a Distinctive Characteristic of Unmethylated Human CpG Island Promoters. *Mol. Cell* **45**, 814-825 (2012).
37. S. R. Hartono, A. Malapert, P. Legros *et al.*, The Affinity of the S9.6 Antibody for Double-Stranded RNAs Impacts the Accurate Mapping of R-Loops in Fission Yeast. *J. Mol. Biol.* **430**, 272-284 (2018).

38. W. Xu, H. Xu, K. Li *et al.*, The R-loop is a common chromatin feature of the Arabidopsis genome. *Nature Plants* **3**, 704-714 (2017).
39. L. A. Sanz, S. R. Hartono, Y. W. Lim *et al.*, Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals. *Mol. Cell* **63**, 167-178 (2016).
40. A. Aguilera, T. Garcia-Muse, R loops: from transcription byproducts to threats to genome stability. *Mol. Cell* **46**, 115-124 (2012).
41. L. Chen, J. Y. Chen, X. Zhang *et al.*, R-ChIP Using Inactive RNase H Reveals Dynamic Coupling of R-loops with Transcriptional Pausing at Gene Promoters. *Molecular Cell* **68**, 745-+ (2017).
42. A. Helmrich, M. Ballarino, L. Tora, Collisions between Replication and Transcription Complexes Cause Common Fragile Site Instability at the Longest Human Genes. *Mol. Cell* **44**, 966-977 (2011).
43. D. Roy, M. R. Lieber, G Clustering Is Important for the Initiation of Transcription-Induced R-Loops In Vitro, whereas High G Density without Clustering Is Sufficient Thereafter. *Mol. Cell Biol.* **29**, 3124-3133 (2009).
44. D. Roy, Z. Zhang, Z. F. Lu, C. L. Hsieh, M. R. Lieber, Competition between the RNA Transcript and the Nontemplate DNA Strand during R-Loop Formation In Vitro: a Nick Can Serve as a Strong R-Loop Initiation Site. *Mol. Cell Biol.* **30**, 146-159 (2010).
45. M. P. Crossley, M. Bocek, K. A. Cimprich, R-Loops as Cellular Regulators and Genomic Threats. *Molecular Cell* **73**, 398-411 (2019).
46. K. F. Yu, F. Chedin, C. L. Hsieh, T. E. Wilson, M. R. Lieber, R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells. *Nat. Immunol.* **4**, 442-451 (2003).
47. M. Jinek, K. Chylinski, I. Fonfara *et al.*, A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816-821 (2012).
48. C. Grunseich, I. X. Wang, J. A. Watts *et al.*, Senataxin Mutation Reveals How R-Loops Promote Transcription by Blocking DNA Methylation at Gene Promoters. *Mol. Cell* **69**, 426-+ (2018).
49. T. Garcia-Muse, A. Aguilera, R Loops: From Physiological to Pathological Roles. *Cell* **179**, 604-618 (2019).
50. R. E. Wellinger, F. Prado, A. Aguilera, Replication fork progression is impaired by transcription in hyperrecombinant yeast cells lacking a functional THO complex. *Mol. Cell Biol.* **26**, 3327-3334 (2006).
51. E. Grabczyk, M. Mancuso, M. C. Sammarco, A persistent RNA.DNA hybrid formed by transcription of the Friedreich ataxia triplet repeat in live bacteria, and by T7 RNAP in vitro. *Nucleic Acids Res.* **35**, 5351-5359 (2007).
52. Y. W. Lim, L. A. Sanz, X. Xu, S. R. Hartono, F. Chedin, Genome-wide DNA hypomethylation and RNA:DNA hybrid accumulation in Aicardi-Goutieres syndrome. *Elife* **4** (2015).
53. V. Bhatia, S. I. Barroso, M. L. Garcia-Rubio *et al.*, BRCA2 prevents R-loop accumulation and associates with TREX-2 mRNA export factor PCID2. *Nature* **511**, 362-365 (2014).
54. S. J. Hill, T. Rolland, G. Adelmant *et al.*, Systematic screening reveals a role for BRCA1 in the response to transcription-associated DNA damage. *Genes Dev* **28**, 1957-1975 (2014).

55. L. Wahba, J. D. Amon, D. Koshland, M. Vuica-Ross, RNase H and multiple RNA biogenesis factors cooperate to prevent RNA:DNA hybrids from generating genome instability. *Mol. Cell* **44**, 978-988 (2011).
56. X. Li, J. L. Manley, Inactivation of the SR protein splicing factor ASF/SF2 results in genomic instability. *Cell* **122**, 365-378 (2005).
57. P. Huertas, A. Aguilera, Cotranscriptionally formed DNA:RNA hybrids mediate transcription elongation impairment and transcription-associated recombination. *Mol. Cell* **12**, 711-721 (2003).
58. E. Y. Chang, C. A. Novoa, M. J. Aristizabal *et al.*, RECQ-like helicases Sgs1 and BLM regulate R-loop-associated genome instability. *J. Cell Biol.* **216**, 3991-4005 (2017).
59. A. Cristini, M. Groh, M. S. Kristiansen, N. Gromak, RNA/DNA Hybrid Interactome Identifies DXH9 as a Molecular Player in Transcriptional Termination and R-Loop-Associated DNA Damage. *Cell Rep.* **23**, 1891-1905 (2018).
60. D. Hodroj, B. Recolin, K. Serhal *et al.*, An ATR-dependent function for the Ddx19 RNA helicase in nuclear R-loop metabolism. *Embo. J.* **36**, 1182-1198 (2017).
61. S. C. Sridhara, S. Carvalho, A. R. Grosso *et al.*, Transcription Dynamics Prevent RNA-Mediated Genomic Instability through SRPK2-Dependent DDX23 Phosphorylation. *Cell Rep.* **18**, 334-343 (2017).
62. I. X. Wang, C. Grunseich, J. Fox *et al.*, Human proteins that interact with RNA/DNA hybrids. *Genome Res.* **28**, 1405-1414 (2018).
63. E. C. Friedberg, A history of the DNA repair and mutagenesis field The discovery of base excision repair. *DNA Repair* **37**, A35-A39 (2016).
64. A. P. Bertolin, S. F. Mansilla, V. Gottifredi, The identification of translesion DNA synthesis regulators: Inhibitors in the spotlight. *DNA Repair (Amst)* **32**, 158-164 (2015).
65. A. R. Lehmann, A. Niimi, T. Ogi *et al.*, Translesion synthesis: Y-family polymerases and the polymerase switch. *DNA Repair (Amst)* **6**, 891-899 (2007).
66. M. F. Goodman, Error-prone repair DNA polymerases in prokaryotes and eukaryotes. *Annu. Rev. Biochem.* **71**, 17-50 (2002).
67. L. S. Waters, B. K. Minesinger, M. E. Wiltrout *et al.*, Eukaryotic Translesion Polymerases and Their Roles and Regulation in DNA Damage Tolerance. *Microbiol. Mol. Biol. R* **73**, 134-+ (2009).
68. E. C. Friedberg, A. R. Lehmann, R. P. P. Fuchs, Trading places: How do DNA polymerases switch during translesion DNA synthesis? (vol 18, pg 499, 2005). *Mol. Cell* **19**, 143-143 (2005).
69. A. R. Lehmann, R. P. Fuchs, Gaps and forks in DNA replication: Rediscovering old models. *DNA Repair (Amst)* **5**, 1495-1498 (2006).
70. S. D. McCulloch, T. A. Kunkel, The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Res.* **18**, 148-161 (2008).
71. A. R. Lehmann, New functions for Y family polymerases. *Mol. Cell* **24**, 493-495 (2006).
72. P. L. Foster, G. Gudmundsson, J. M. Trimarchi, H. Cai, M. F. Goodman, Proofreading-defective DNA polymerase II increases adaptive mutation in Escherichia coli. *Proc Natl Acad Sci U S A* **92**, 7951-7955 (1995).
73. S. Rangarajan, G. Gudmundsson, Z. Qiu, P. L. Foster, M. F. Goodman, Escherichia coli DNA polymerase II catalyzes chromosomal and episomal DNA synthesis in vivo. *Proc Natl Acad Sci U S A* **94**, 946-951 (1997).

74. R. L. Frisch, Y. Su, P. C. Thornton *et al.*, Separate DNA Pol II- and Pol IV-dependent pathways of stress-induced mutation during double-strand-break repair in Escherichia coli are controlled by RpoS. *J. Bacteriol.* **192**, 4694-4700 (2010).
75. C. J. Kenyon, G. C. Walker, DNA-Damaging Agents Stimulate Gene-Expression at Specific Loci in Escherichia-Coli. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences* **77**, 2819-2823 (1980).
76. S. Mead, A. Vaisman, M. Valjavec-Gratian *et al.*, Characterization of polVR391: a Y-family polymerase encoded by rumA'B from the IncJ conjugative transposon, R391. *Mol. Microbiol.* **63**, 797-810 (2007).
77. M. F. Goodman, R. Woodgate, Translesion DNA polymerases. *Cold Spring Harb Perspect Biol* **5**, a010363 (2013).
78. R. E. Johnson, S. Prakash, L. Prakash, Efficient bypass of a thymine-thymine dimer by yeast DNA polymerase, Poleta. *Science* **283**, 1001-1004 (1999).
79. J. H. Yoon, L. Prakash, S. Prakash, Highly error-free role of DNA polymerase eta in the replicative bypass of UV-induced pyrimidine dimers in mouse and human cells. *Proc Natl Acad Sci U S A* **106**, 18219-18224 (2009).
80. J. E. Sale, A. R. Lehmann, R. Woodgate, Y-family DNA polymerases and their role in tolerance of cellular DNA damage. *Nat. Rev. Mol. Cell Bio.* **13**, 141-152 (2012).
81. B. F. Yuan, C. J. You, N. Andersen *et al.*, The Roles of DNA Polymerases kappa and iota in the Error-free Bypass of N-2-Carboxyalkyl-2 '-deoxyguanosine Lesions in Mammalian Cells. *J. Biol. Chem.* **286**, 17503-17511 (2011).
82. D. F. Jarosz, V. G. Godoy, J. C. Delaney, J. M. Essigmann, G. C. Walker, A single amino acid governs enhanced activity of DinB DNA polymerases on damaged templates. *Nature* **439**, 225-228 (2006).
83. J. A. Brown, J. D. Fowler, Z. C. Suo, Kinetic Basis of Nucleotide Selection Employed by a Protein Template-Dependent DNA Polymerase. *Biochemistry-us.* **49**, 5504-5510 (2010).
84. E. Mentegari, E. Crespan, L. Bavagnoli *et al.*, Ribonucleotide incorporation by human DNA polymerase eta impacts translesion synthesis and RNase H2 activity. *Nucleic Acids Res.* **45**, 2600-2614 (2017).
85. P. Sarkies, C. Reams, L. J. Simpson, J. E. Sale, Epigenetic Instability due to Defective Replication of Structured DNA. *Mol. Cell* **40**, 703-713 (2010).
86. C. M. Wickramasinghe, H. Arzouk, A. Frey, A. Maiter, J. E. Sale, Contributions of the specialised DNA polymerases to replication of structured DNA. *DNA Repair* **29**, 83-90 (2015).
87. E. Walsh, X. Wang, M. Y. Lee, K. A. Eckert, Mechanism of replicative DNA polymerase delta pausing and a potential role for DNA polymerase kappa in common fragile site replication. *J. Mol. Biol.* **425**, 232-243 (2013).
88. K. L. Burr, S. Velasco-Miguel, V. S. Duvvuri *et al.*, Elevated mutation rates in the germline of Polkappa mutant male mice. *DNA Repair (Amst)* **5**, 860-862 (2006).
89. R. Betous, L. Rey, G. L. Wang *et al.*, Role of TLS DNA Polymerases eta and kappa in Processing Naturally Occurring Structured DNA in Human Cells. *Mol. Carcinogen.* **48**, 369-378 (2009).
90. J. C. Delaney, J. M. Essigmann, Assays for determining lesion bypass efficiency and mutagenicity of site-specific DNA lesions in vivo. *DNA Repair, Pt A* **408**, 1-15 (2006).
91. B. F. Yuan, Y. S. Wang, Mutagenic and cytotoxic properties of 6-thioguanine, S-6-methylthioguanine, and guanine-S-6-sulfonic acid. *J. Biol. Chem.* **283**, 23665-23670 (2008).

92. C. J. You, Y. S. Wang, Mass Spectrometry-Based Quantitative Strategies for Assessing the Biological Consequences and Repair of DNA Adducts. *Accounts. Chem. Res.* **49**, 205-213 (2016).
93. C. J. You, Y. S. Wang, Quantitative measurement of transcriptional inhibition and mutagenesis induced by site-specifically incorporated DNA lesions in vitro and in vivo. *Nature Protocols* **10**, 1389-1406 (2015).
94. M. R. Wilkins, J. C. Sanchez, A. A. Gooley *et al.*, Progress with proteome projects: Why all proteins expressed by a genome should be identified and how to do it. *Biotechnology and Genetic Engineering Reviews, Vol 13* **13**, 19-50 (1996).
95. J. Parsons, C. Francavilla, 'Omics Approaches to Explore the Breast Cancer Landscape. *Frontiers in Cell and Developmental Biology* **7** (2020).
96. M. F. Tong, H. M. Liu, J. Y. Hao, D. M. Fan, Comparative pharmacoproteomics reveals potential targets for berberine, a promising therapy for colorectal cancer. *Biochem. Bioph. Res. Co* **525**, 244-250 (2020).
97. S. Saleem, S. Tariq, I. Aleem *et al.*, Proteomics analysis of colon cancer progression. *Clinical Proteomics* **16** (2019).
98. J. C. Rogers, R. D. Bomgarden, Sample Preparation for Mass Spectrometry-Based Proteomics; from Proteomes to Peptides. *Modern Proteomics - Sample Preparation, Analysis and Practical Applications* **919**, 43-62 (2016).
99. E. J. Dupree, M. Jayathirtha, H. Yorkey *et al.*, A Critical Review of Bottom-Up Proteomics: The Good, the Bad, and the Future of This Field. *Proteomes* **8** (2020).
100. J. D. Martell, T. J. Deerinck, Y. Sancak *et al.*, Engineered ascorbate peroxidase as a genetically encoded reporter for electron microscopy. *Nat. Biotechnol.* **30**, 1143-+ (2012).
101. H. W. Rhee, P. Zou, N. D. Udeshi *et al.*, Proteomic Mapping of Mitochondria in Living Cells via Spatially Restricted Enzymatic Tagging. *Science* **339**, 1328-1331 (2013).
102. E. Choi-Rhee, H. Schulman, J. E. Cronan, Promiscuous protein biotinylation by Escherichia coli biotin protein ligase. *Protein Sci.* **13**, 3043-3050 (2004).
103. A. L. Richards, M. Eckhardt, N. J. Krogan, Mass spectrometry-based protein-protein interaction networks for the study of human diseases. *Molecular Systems Biology* **17** (2021).
104. I. F. Chang, Mass spectrometry-based proteomic analysis of the epitope-tag affinity purified protein complexes in eukaryotes. *Proteomics* **6**, 6158-6166 (2006).
105. F. A. Ran, P. D. Hsu, J. Wright *et al.*, Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* **8**, 2281-2308 (2013).
106. W. Jiang, D. Bikard, D. Cox, F. Zhang, L. A. Marraffini, RNA-guided editing of bacterial genomes using CRISPR-Cas systems. *Nat. Biotechnol.* **31**, 233-239 (2013).
107. E. L. Huttlin, L. Ting, R. J. Bruckner *et al.*, The BioPlex Network: A Systematic Exploration of the Human Interactome. *Cell* **162**, 425-440 (2015).
108. E. L. Huttlin, R. J. Bruckner, J. A. Paulo *et al.*, Architecture of the human interactome defines protein communities and disease networks. *Nature* **545**, 505-+ (2017).
109. M. Dalvai, J. Loehr, K. Jacquet *et al.*, A Scalable Genome-Editing-Based Approach for Mapping Multiprotein Complexes in Human Cells. *Cell Rep.* **13**, 621-633 (2015).
110. A. L. Richards, M. Eckhardt, N. J. Krogan, Mass spectrometry-based protein-protein interaction networks for the study of human diseases. *Mol Syst Biol* **17**, e8792 (2021).
111. L. M. Smith, J. Z. Sanders, R. J. Kaiser *et al.*, Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674-679 (1986).

112. Z. G. Chidgeavadze, R. S. Beabealashvilli, A. M. Atrazhev *et al.*, 2',3'-Dideoxy-3' aminonucleoside 5'-triphosphates are the terminators of DNA synthesis catalyzed by DNA polymerases. *Nucleic Acids Res.* **12**, 1671-1686 (1984).
113. J. Shendure, H. Ji, Next-generation DNA sequencing. *Nat. Biotechnol.* **26**, 1135-1145 (2008).
114. P. Wulfridge, K. Sarma, A nuclease- and bisulfite-based strategy captures strand-specific R-loops genome-wide. *Elife* **10** (2021).
115. J. Nadel, R. Athanasiadou, C. Lemetre *et al.*, RNA:DNA hybrids in the human genome have distinctive nucleotide characteristics, chromatin composition, and transcriptional relationships. *Epigenetics & Chromatin* **8** (2015).
116. L. A. Sanz, S. R. Hartono, Y. W. Lim *et al.*, Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals. *Mol. Cell* **63**, 167-178 (2016).
117. K. Wang, H. H. Wang, C. H. Li *et al.*, Genomic profiling of native R loops with a DNA-RNA hybrid recognition sensor. *Science Advances* **7** (2021).
118. J. G. Dumelie, S. R. Jaffrey, Defining the location of promoter associated R-loops at near-nucleotide resolution using bisDRIP-seq. *Elife* **6** (2017).
119. T. Wu, R. T. Lyu, Q. C. You, C. He, Kethoxal-assisted single-stranded DNA sequencing captures global transcription dynamics and enhancer activity in situ (vol 17, pg 515, 2020). *Nature Methods* **17**, 749-749 (2020).
120. X. C. Weng, J. Gong, Y. Chen *et al.*, Keth-seq for transcriptome-wide RNA structure mapping. *Nat. Chem. Biol.* **16**, 489-+ (2020).
121. T. Wu, R. Lyu, C. He, spKAS-seq reveals R-loop dynamics using low-input materials by detecting single-stranded DNA with strand specificity. *Science Advances* **8** (2022).

Chapter 3

Polymerase κ Recruits DDX23 to Promote R-loop Resolution

Introduction

Endogenous and exogenous sources of damaging agents can attack genomic DNA to yield numerous DNA damage products (1). The ensuing DNA lesions can stall replication forks and induce mutations, thereby leading to accelerated aging, cancer, and neurodegeneration (2). Translesion synthesis (TLS) is a DNA damage tolerance mechanism that cells employ to cope with unrepaired DNA lesions (3). TLS is mainly carried out by Y-family DNA polymerases, which are found in all domains of life and have functions in bypassing various types of DNA lesions (4-7).

Among the Y-family DNA polymerases, Pol κ is highly conserved in all three kingdoms of life, including archaea, bacteria, and eukaryotes (8, 9). While the basis for this marked conservation is unclear, previous studies suggest that this may reside in its ability in bypassing accurately and efficiently endogenously induced minor-groove N2-dG adducts (10, 11). In this vein, biochemical experiments showed that Pol κ catalyzes accurate nucleotide insertion opposite minor-groove N2-dG adducts (10, 12-14). Genetic ablation of Pol κ led to increased frequencies of mutations induced by N2-dG adduct of benzo[a]pyrene diolepoxide (BPDE) (15, 16) and mutagenic bypass of other N2-alkylated dG lesions in mammalian cells (11, 17). Pol κ also supports DNA replication past abasic sites, thymidine glycol, DNA-peptide cross-links, as well as intra- and interstrand DNA cross-links induced by cisplatin and mitomycin C, and the polymerase is also engaged in the repair of DNA interstrand cross-links (18-24). Additionally, Pol κ assumes important functions in DNA syntheses across non-B form DNA, such as common fragile sites

(25), tandem repeat minisatellites (26), and guanine quadruplexes (G4s) (27). Moreover, Pol κ's role in protecting against DNA damage can be independent of its catalytic function (28), where Pol κ was found to be involved in DNA repair synthesis during nucleotide excision repair (NER) (29).

Given the diverse roles of Pol κ, we reason that a proteome-wide assessment of Pol κ-interacting proteins will improve our understanding of the functions and regulations of Pol κ. In this study, we provided the first unbiased exploration of Pol κ interactome using two independent proteomic strategies, and our results revealed a number of candidate interaction partners of Pol κ, including DDX23. We also explored the genome-wide occupancy of Pol κ in chromatin using ChIP-Seq analysis, which uncovered the enrichment of Pol κ at R-loop sites. Moreover, we found that Pol κ promotes the recruitment of DDX23 to R-loop loci in chromatin and facilitates DDX23-mediated resolution of R-loop structures.

Materials and Methods

Cell lines

HEK293T and U2OS cells were purchased from ATCC. All cell lines used in this study were tested to be free of mycoplasma contamination using LookOut Mycoplasma PCR Kit (Sigma, MP0035). Cells were maintained in DMEM (Life Technologies) supplemented with 10% FBS (Invitrogen) and 1% penicillin/streptomycin (v/v) at 37°C in a humidified incubator with 5% CO2. For genomic DNA extraction, cells were harvested at 80% confluency. Pol κ knockout cells in HEK293T background were generated previously using the CRISPR-Cas9 method (30). DDX23 knockout cells were generated in the HEK293T background using CRISPR-Cas9, following previously reported procedures (31). The guide RNAs were designed according to a previously described method (32), and their sequences are listed in Table 3-1.

Targeted integration of Flag tag using CRISPR-Cas9

Genome editing-based integration of three tandem repeats of the Flag epitope tag (3× Flag tag) into the C-terminus of endogenous Pol κ and DDX23 was conducted following the previously reported procedures (33). DNA sequences for the production of sgRNA targeting the POLK and DDX23 genes were inserted into the hSpCas9 plasmid pX330 (Addgene, Cambridge, MA, USA). The template donor plasmids for tagging POLK and DDX23 were synthesized (gBlock, IDT) and inserted into pUC19. The constructed Cas9 plasmid and the donor plasmids were transfected into HEK293T cells using Lipofectamine 2000 (Invitrogen, Carlsbad, CA, USA). At 24 h following transfection, the cells were diluted and seeded into a 96-well plate at a concentration of approximately 0.5 cell per well. After 2 weeks, single colonies were transferred to a 24-well plate. Genomic DNA was extracted for PCR amplification and agarose gel electrophoresis to screen for clones with successful insertion of the Flag tag. Sanger sequencing and Western blot were performed to confirm the correct insertion. The guide RNAs were designed according to a previously described method (32), and their sequences are listed in Table 3-1.

APEX-labeling experiments

The human POLK gene was amplified from a cDNA library of HEK293T cells to introduce a 5′ Not1 site and a 3′ BamHI site and subcloned into a mito-V5-APEX2 vector to replace the mitochondria-targeting sequence with the coding sequence of human Pol κ (34). The APEX2-NLS vector was a gift from Prof. Alice Ting (35). APEX-labeling experiments were conducted following the previously published procedures with minor modifications (36). The constructed APEX2 plasmids were transfected into HEK293T cells using Lipofectamine 2000. At 24 h following transfection, the cells were incubated in complete medium with 0.5 mM biotin-phenol at 37°C for 30 min. Hydrogen peroxide was subsequently added to a final concentration of 1 mM,

and the cells were incubated at room temperature for 1 min. The labeling medium was quickly aspirated off, and the cells were washed five times with a quenching solution, which contained PBS supplemented with 10 mM sodium ascorbate, 10 mM sodium azide, and 5 mM Trolox. For LC-MS/MS and Western blot analyses, CelLytic™ M (sigma) buffer supplemented with 2 M urea, 1× protease inhibitor cocktail, 1 mM PMSF and quenchers (10 mM sodium ascorbate, 10 mM sodium azide, and 5 mM Trolox) was immediately added to the cell pellet. After resuspension, the sample was placed on the rotator at 4°C for 30 min. The cell lysate was cleared by centrifugation at 16,000 g for 15 min. The protein concentrations were measured using Bradford Quick Start Protein Assay kit (Bio-Rad), and the lysate was used immediately for further analyses or stored at -80°C. One mg of whole-cell lysate was used for the subsequent anti-biotin beads pull-down and LC-MS/MS analysis.

For removing excess free biotin-phenol, the collected supernatant was subjected to buffer exchange to 8 M urea, 50 mM NH_4HCO_3 with a Amicon® Ultra-4 Centrifugal Filter Unit (Millipore Sigma, 10 kDa). Dithiothreitol (DTT, Sigma) was added to a 20 mM final concentration and the solution was incubated at 37°C for 1 h. Each sample was treated with 55 mM iodoacetamide (IAA, Sigma) for 30 min in the dark. The resulting sample was subjected to buffer exchange with 50 mM NH_4HCO_3 three times. Trypsin (Thermo Scientific) was added to the mixture at a 1:100 enzyme:substrate ratio and the samples were incubated at 37°C overnight. After digestion, the sample was centrifuged and the flow-through was collected. Additional aliquot of 50 mM NH_4HCO_3 (500 µL) was used to rinse the tube. The digested samples were dried using a Speedvac. Peptides were dissolved in 1 mL IAP buffer containing 50 mM MOPS, 10 mM Na_2HPO_4, 50 mM NaCl at pH 7.5. The digested peptides were incubated with 100 µg pre-washed Anti-biotin beads (Immune Chem Pharmaceuticals) at 4°C overnight. The beads were washed twice with IAP buffer and then twice with H_2O. The biotinylated peptides were eluted

from the beads using 0.15% TFA in a Thermomixer at 60°C for 10 min with mixing at 1,200 rpm. The elution was repeated twice, and the supernatants were combined for desalting using OMIX C18 Tips (Agilent), and analyzed by LC-MS/MS.

Affinity pull-down of Pol κ-binding proteins

For pull-down assay, Pol κ-Flag and parental HEK293T cells were collected and lysed in CelLytic M cell lysis reagent (Sigma) with 1x protease inhibitor cocktail (Sigma). After lysis, the cell lysate was incubated with 30 μL pre-equilibrated anti-Flag M2 affinity gel (Sigma) at 4°C overnight. The beads were washed five times with a washing buffer containing 50 mM Tris (pH 8.0), 150 mM NaCl, and 0.1% Tween 20 (v/v). The bound proteins were subsequently eluted by boiling for 5 min in 2x protein loading buffer (Bio-rad). The resulting mixture was centrifuged, and the supernatant was loaded onto a 15% SDS-PAGE gel, which was run at 120 V for a very short time (~10 min). The gel was washed to remove the running buffer and stained with Coomassie blue (VWR) for 0.5 h. The gel was subsequently destained overnight with a destaining solution (ethanol: H_2O: acetic acid, 4.5:4.5:1, v/v). Gel slices containing the proteins were excised and cut into 1 mm^2 cubes.

The eluted proteins were digested in-gel following a previously described protocol (33). Briefly, excess SDS in the gel was removed by incubating with 500 μL of 100 mM NH_4HCO_3/CH_3CN (1:1, v/v) in a thermomixer at 37°C with interval mixing at 1,200 rpm. The supernatant was removed and the gel pieces were dehydrated with acetonitrile for 5 min. Gel pieces were further dehydrated in a vacuum centrifuge for 1-2 min. Proteins were then reduced with 10 mM dithiothreitol (DTT, Sigma) in 25 mM NH_4HCO_3 at 55°C for 1 h, and subsequently alkylated by incubating with 55 mM iodoacetamide (IAA, Sigma) in the dark for 1 h. Gel pieces were dehydrated with acetonitrile and washed three times with 1 mL of 25 mM NH_4HCO_3. Proteins

were then digested with trypsin at 37°C overnight. After digestion, the peptides were eluted from the gel by incubating, with vigorous shaking at 37°C for 15 min, first in 200 µL 5% formic acid (v/v) in 25 mM NH4HCO3, then in 200 µL 5% formic acid in 25 mM NH_4HCO_3 and 50% acetonitrile (v/v), and finally in 200 µL 5% acetic acid in 25 mM NH_4HCO_3 and 70% acetonitrile (v/v). After elution, the peptide fractions were combined, evaporated to dryness, desalted using OMIX C18 Tips (Agilent), and analyzed by LC-MS/MS.

LC-MS/MS

LC-MS/MS experiments were performed as previously described with minor modifications (37). Briefly, the affinity pull-down samples were analyzed using an EASY-nLC 1200 system coupled with a Q Exactive Plus quadrupole-Orbitrap mass spectrometer (Thermo Fisher Scientific). The HPLC separation was conducted using a trapping column followed by a separation column, both packed in-house with ReproSil-Pur C18-AQ resin (3 µm, Dr. Maisch HPLC GmbH, Germany). The peptides were separated using a 200-min linear gradient of 2-40% acetonitrile in 0.1% formic acid at a flow rate of 300 nL/min. The mass spectrometer was set up in the positive-ion mode, and the spray voltage was 1.8 kV. MS/MS were recorded in a data-dependent acquisition mode, in which one full-scan MS was followed with 25 MS/MS scans. An Orbitrap Fusion Lumos Tribrid mass spectrometer (Thermo Fisher Scientific) coupled to a nanoelectrospray ion source (NanoFlex, Thermo Fisher Scientific) was used for analyzing APEX labeling samples in the data-dependent acquisition (DDA) mode.

Data Analysis

LC-MS/MS data were processed using MaxQuant version 1.6.14 with the default parameters unless otherwise specified (http://www.maxquant.org) (38). Database search was performed using the Andromeda search engine included with MaxQuant with the UniProt human sequence

database (UP000005640). The mass tolerances for precursor and fragment ions were set at 4.5 ppm and 20 ppm, respectively. Digestion enzyme specificity was set to trypsin with a maximum of 2 missed cleavages. A minimum peptide length of 7 residues was required for protein identification. N-terminal acetylation and methionine oxidation were set as variable modifications, and cysteine carboamidomethylation was set as a fixed modification. For identifying biotinylated peptides, biotin-phenol modifications (+361.1460 Da) of tyrosine, cysteine, tryptophan and histidine were set as variable modifications. “Match between runs” based on accurate m/z and retention time was enabled with a 5-min alignment time window. Label-free quantitation (LFQ) was performed using the MaxLFQ algorithm in MaxQuant (39). Protein LFQ intensities were calculated from the median of pairwise intensity ratios of peptides identified in two or more samples and adjusted to the cumulative intensity across samples. Quantification was performed using razor and unique peptides, including those modified by acetylation (protein N-terminus) and oxidation (Met). A minimum peptide ratio of 1 was required for protein intensity normalization.

Data processing was performed using Perseus version 1.6.13.0 (http://www.perseus-framework.org). For the APEX labeling samples, intensity of biotinylated peptides was used for quantification. For the affinity pull-down results, contaminants and protein groups identified by a single peptide were filtered out from the data set. Protein group LFQ intensities were log2-transformed to reduce the effect of outliers. For statistical comparisons between proteomes, protein groups with missing LFQ values were assigned values using imputation. Missing values were assumed to be biased toward low-abundance proteins that were below the MS detection limit, referred to as “missing not at random”, an assumption that is frequently made in proteomics studies. The missing values were replaced with random values taken from a median downshifted Gaussian distribution to simulate low-abundance LFQ values. Imputation was performed

separately for each sample from a distribution with a width of 0.3 and a downshift of 1.8. Differences in log2(LFQ intensity) were calculated between experimental and control groups. A two-tailed, unpaired, Student's t-test was employed to identify differentially enriched proteins. Visualization of the results was performed with volcano plots and Venn diagrams using the R libraries ggplot2 and VennDiagram.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE (40) partner repository with the dataset identifier PXD038719.

Pull-down and Western blot

Protein pull-down was performed with anti-Flag M2 affinity gel following the aforementioned procedures. Plasmids for expressing truncated DDX23 proteins were kindly provided by Dr. Shaochun Yuan (41). Plasmids for expressing EGFP-tagged wild-type and mutant Pol κ (eGFP-C1-Pol κ-WT, eGFP-C1-Pol κ-CD, and eGFP-C1-Pol κ-UBZ) were kindly provided by Dr. Tony T. Huang (42). Western blot analysis was conducted following previously published procedures with minor modifications (31), where protein samples were separated on a SDS-PAGE gel and transferred to a nitrocellulose membrane (Bio-Rad). After blocking with a blotting-grade blocker (Bio-Rad), the membrane was incubated in a solution containing primary antibody and 5% BSA (w/v) at 4°C overnight, and then incubated with the HRP-conjugated secondary antibody in a 5% blotting-grade blocker (w/v) at room temperature for 1 h. The immunoblots were detected using ECL Western blotting detection reagent (Amersham). Primary antibodies used in this study included anti-Flag (Cell Signaling Technology, 2368S; 1:2000), anti-HA (Cell Signaling Technology, 3724S; 1:2000), anti-GFP (Proteintech, 50430-2-AP; 1:5000), anti-GAPDH (Santa Cruz, sc-32233; 1:5000), anti-Tubulin (Santa Cruz, SC-32293; 1:5000), anti-DDX23 (Proteintech, 10199-2-AP; 1:2000), and anti-Pol κ (Abclonal, A6122; 1:2000).

Chromatin immunoprecipitation (ChIP) and next-generation sequencing

Chromatin immunoprecipitation of Pol κ was performed as previously described with a few modifications (33). Briefly, 2×107 cells were cross-linked by adding formaldehyde dropwise directly to the media for a final concentration of 1% (w/v) and rotating gently at room temperature for 10 min. The reaction was quenched by adding 2.5 M glycine to the medium until its final concentration reached 125 mM, and the mixture was incubated at room temperature with shaking for 5 min. After washing with 1× PBS three times, the cells were resuspended in PBS. After centrifugation, the cell pellet was resuspended in RIPA buffer [50 mM Tris-HCl, pH 8.0, 150 mM NaCl, 2 mM EDTA, 1% NP40 (v/v), 0.5% sodium deoxycholate (w/v), 0.1% SDS (w/v) and 1x protease inhibitor cocktail] at 4 °C for 30 min with rotation. Chromatin was sheared using a Covaris S220 sonicator at 4°C for 10 min with a peak incident power of 140, a duty cycle of 10%, and 200 cycles per burst. After centrifugation at 16,000g for 15 min, the supernatant was incubated with pre-blocked anti-Flag M2 affinity gel (Sigma) on a rotator at 4°C overnight. After washing, DNA was eluted from the beads with 120 µL 100 mM $NaHCO_3$ and 1% SDS (w/v) at 68°C for 2 h. Cross-links were subsequently reversed by adding 4.8 µL of 5 M NaCl and 2 µL RNase A (10 mg/mL) and incubated with shaking at 65°C overnight. Proteins in the resulting DNA samples were removed with 2 µL proteinase K (20 mg/mL) and incubated with shaking at 60°C for 2 h. Finally, the DNA was purified using Cycle Pure Kit (Omega).

Purified DNA was quantified and verified on an Agilent 2100 Bioanalyzer. The library was then constructed using NEBNext Ultra DNA Library Prep Kit (NEB, E7103S) following the manufacturer's instructions (43). Subsequently, the purified DNA libraries were assessed using an Agilent 2100 Bioanalyzer and multiplexed for sequencing on a HiSeq4000 system (Illumina).

The sequencing reads of ChIP-seq were checked with FastQC and aligned to the human hg19 reference genome using Bowtie2 with the default configuration (44). After alignment, the sequencing duplicates were removed using Picard MarkDuplicates (http://broadinstitute.github.io/picard/). For genome-wide identification of Pol κ-binding sites, peak calling was performed using the model-based analysis of ChIP-seq (MACS2) (45) with the following configuration: macs2 callpeak -t treat.bam -c ctrl.bam -f BAM -g 2.7e+9 -n –ourdir macs2. Integrative Genomics Viewer (46) was used to visualize the mapping results. Overlap of two biological replicates was defined as high-confidence peaks, which were employed in further analysis. Genomic annotations were performed using Homer (47). The intersection between bed files was performed using BEDTools (48). Plotprofile was performed using deepTools (49). The ChIP-Seq data generated in this study have been deposited into the NCBI Gene Expression Omnibus (GEO) under accession number GSE220603.

Chromatin immunoprecipitation-quantitative PCR analysis of DDX23

Crosslinking, nuclei isolation, and sonication were performed with 8 × 107 cells following the aforementioned procedures (50). Immunoprecipitation was conducted following previously published procedures with minor modifications. Sheared chromatin was incubated with 60 μL of blocked anti-Flag M2 affinity gel overnight with rotation at 4°C. The beads were washed with cold PBS containing 1× protease inhibitor cocktail (Sigma). Complexes were eluted with 500 μL of PBS (with protease inhibitor cocktail) supplemented with 150 ng/μL 3× Flag peptide (Sigma) and rotated at 4°C for 4 h. After centrifugation at 8,000g for 1 min, the supernatant was collected and incubated with 20 μL prewashed Strep-Tactin Sepharose beads (Fisher) overnight with rotation at 4°C. After washing with cold PBS, DNA was eluted with 120 μL of elution buffer [100 mM NaHCO3 and 1% SDS (w/v)] supplemented with 2.5 mM biotin, incubated on a thermomixer at 600 rpm for 1 h at 65°C. Cross-links were subsequently reversed by adding 4.8

μL of 5 M NaCl and 2 μL RNase A, and incubating with shaking at 65°C overnight. Proteins in the resulting DNA samples were removed by treating with proteinase K (2 μL, 20 mg/mL) at 60°C with shaking for 2 h. Finally, the DNA was purified using Cycle Pure Kit (Omega). Quantitative PCR was performed using standard protocols with Luna® Universal qPCR Master Mix (NEB, M3003X). Analysis was carried out using the ΔΔCt method (fold enrichment), as described previously (31).

Establishment of a stable cell line expressing catalytically inactive RNase H1

The cells with stable expression of the RNase H1-D210N or WKKD mutant protein were generated by following previously published procedures (51). Briefly, 2-3×105 HEK293T cells were seeded in a 6-well plate. When cells reached 60-70% confluence, 1.5 μg of pPyCAG-RNASEH1 (D210N/WKKD) plasmid was transfected into cells using Lipofectamine 2000. Two days after transfection, hygromycin B was added to a final concentration of 100-200 μg/mL. The medium was replaced with fresh medium supplemented with hygromycin B in every 2-3 days until new clones repopulate. The selection process took around 2 weeks.

R-ChIP

Chromatin immunoprecipitation of R-loops was performed as previously described with a few modifications (51). Briefly, 2×107 cells were cross-linked with formaldehyde by following the aforementioned procedures. After washing with PBS twice, the cell pellet was resuspended in 1 mL lysis buffer [10 mM Tris-Cl, pH 8.0, 10 mM NaCl, and Igepal CA-630 (0.5%, v/v)] with 1× protease inhibitor cocktail and 1 μL RiboLock RNase inhibitor (Thermo Fisher) at 4°C for 30 min with rotation. The lysate was centrifuged at 4°C to remove the supernatant. The nuclear pellet was washed once with 1 mL of cell lysis buffer and resuspended in 1 mL of nuclear lysis buffer [50 mM Tris-Cl, pH 8.0, 10 mM EDTA, pH 8.0, and 1% (w/v) SDS] with 1× protease inhibitor

cocktail and RiboLock RNase inhibitor. Chromatin was sheared using a Covaris S220 sonicator at 4°C for 10 min with a peak incident power of 140, a duty cycle of 10%, and 200 cycles per burst. The samples were kept on ice for 30 min until the SDS was precipitated. After centrifugation at 16,000 g for 15 min, the supernatant containing sheared chromatin was transferred to a new 1.5-mL LoBind tube, and the following reagents were added to bring the volume to 1.3 mL: 144 μL of TE buffer, 130 μL of 10% (v/v) Triton X-100, 13 μL of 10% (w/v) sodium deoxycholate, 13 μL of 100× proteinase inhibitor cocktail and 0.5 μL of RiboLock RNase inhibitor. Chromatin suspension was incubated overnight with the V5-antibody-conjugated beads on a tube roller at 4°C. The beads were washed successively twice with 1 mL wash buffer I [20 mM Tris-Cl, pH 8.0, 0.2 mM EDTA, pH 8.0, 1% (v/v) Triton X-100, 0.1% (w/v) SDS and 150 mM NaCl], followed by 1 mL of wash buffer II [20 MM Tris-Cl, pH 8.0, 0.2 mM EDTA, pH 8.0, 1% (v/v) Triton X-100, 0.1% (w/v) SDS and 500 mM NaCl], 1 mL of wash buffer III [10 mM Tris-Cl, pH 8.0, 0.1 mM EDTA, pH 8.0, 1% (v/v) Igepal CA-630, 250 mM LiCl 1% (w/v) sodium deoxycholate] and 1 mL of TE buffer. After washing, DNA was eluted from the beads with 170 μL of elution buffer [10 mM Tris-Cl, pH 8.0, 10 mM EDTA, pH 8.0, and 1% (w/v) SDS] for 30 min at 65°C on a thermomixer. Cross-links were subsequently reversed by adding 4.8 μL of 5 M NaCl and 2 μL RNase A (10 mg/mL) and incubated with shaking at 65°C overnight. Proteins in the resulting DNA samples were removed with 2 μL proteinase K (20 mg/mL) and incubated with shaking at 60°C for 2 h. Finally, the DNA was purified using Cycle Pure Kit (Omega).

Purified DNA was quantified and verified on an Agilent 2100 Bioanalyzer. The library was then constructed using NEBNext Ultra DNA Library Prep Kit (NEB, E7103S) following the manufacturer's instructions (43). Subsequently, the purified DNA libraries were assessed using an Agilent 2100 Bioanalyzer and multiplexed for sequencing on a MGISEQ-2000 platform (BGI). The data were processed following the above-mentioned procedures.

Expression and purification of recombinant GFP-RNase H1

The plasmid for expressing N-terminal tandem His-GFP-tagged RNase H1 D210N (GFP-dRNH1) was a gift from Dr. Karlene Cimprich (52). Rosetta (DE3) *E.* coli cells were transformed with the GFP-dRNH1 plasmid and grown on LB agar with 50 μg/mL kanamycin at 37°C overnight. A single colony was amplified in a 5-mL LB medium containing 50 μg/mL kanamycin at 37°C overnight and then diluted to a 500 mL LB medium. After the optical density at 600 nm reached approximately 0.6, the culture was cooled to 16°C, and the protein expression was induced with 0.3 mM IPTG (Sigma) at 16°C for approximately 20 h. The cells were then collected by centrifugation and lysed at 4°C by sonication in a 15 mL 1× PBS buffer (Fisher BioReagents) containing 1 mM phenylmethylsulfonyl fluoride. The suspension was centrifuged at 14,000 g for 15 min, and the supernatant was collected and purified with HisTrap™ Fast Flow column (Cytiva). The column was washed with 1 mL 1× PBS buffer containing 40 mM imidazole and 1 mM PMSF at 1 mL/min until the flow-through became colorless. Samples were gradient eluted with 1× PBS buffer containing 50-500 mM imidazole and 1 mM PMSF at 0.5 mL/min until the eluent became colorless. Peak fractions containing GFP-dRNH1 were pooled and dialyzed overnight against a dialysis buffer containing 0.2 × PBS, 1.0 mM DTT and 1.0 mM PMSF. Dialyzed samples were applied to a HiTrap Heparin HP affinity column (Cytiva), and the column was washed with 1 mL 0.2 × PBS for 12 times at 1 mL/min. The GFP-dRNH1 was eluted with 0.2 × PBS containing 1.5 M NaCl until the eluent became colorless. Fractions containing GFP-dRNH1 were combined and dialyzed overnight into a storage buffer (0.2 × PBS containing 1 mM DTT and 1 mM PMSF). Purified GFP-RNH1 was analyzed by Coomassie-stained SDS-PAGE with BSA as standard to determine protein purity and concentrations. The stock solution of dialyzed GFP-RNH1 at 1 mg/mL was aliquoted, flash-frozen, stored at −80°C and freshly thawed for each experiment.

Fluorescence for monitoring R-loops

HEK293T cells and isogenic cells deficient in DDX23 or Pol κ grown on coverslips were fixed with chilled methanol at room temperature for 15 min. The cells were then washed with PBS-TX (PBS containing 0.1% Triton X-100) and blocked with antibody dilution solution (AbDil-Tx, PBS containing 2% BSA, 0.05% NaN_3) at 4°C for 6 h. Samples were then incubated with a 1:6,000 dilution of GFP-dRNH1 at 1 mg/mL in antibody dilution buffer for 1.5 h at 37°C. After washing with PBS-TX three times for 5 min, the samples were incubated with 4'-6-diamidino-2-phenylindole (DAPI) (Sigma) for 5 min to stain the DNA. Samples were then washed three times with PBS-TX, and coverslips were mounted onto glass slides using ProLong™ Diamond Antifade Mountant (Invitrogen) and stored at 4°C in the dark. Images were taken using a Zeiss 880 Inverted (Carl Zeiss, Oberkochen, Germany) confocal microscope with a 40x/1.4 oil immersion and quantified with ImageJ.

Co-localization analysis

The pRK7-Flag-DDX23 was constructed by inserting the coding sequence of DDX23 into the pRK7-Flag plasmids following our previously published procedures (53). Catalytically inactive DDX23 (A435V and K441N) plasmid was generated using tandem site-mutagenesis according to a previous report (54). pRK7-Flag-POLK-WT, pRK7-Flag-POLK-CD, and pRK7-Flag-POLK-UBZ were constructed by PCR amplification from the corresponding EGFP-tagged vectors and inserted into the pRK7 backbone. The constructed Flag-tagged plasmids were transfected into U2OS cells using Lipofectamine 2000. At 24 h following transfection, the cells were fixed with chilled methanol at room temperature for 15 min. The cells were then washed with PBS-TX and blocked with antibody dilution solution at 4°C for 6 h. Samples were then incubated with a 1:250 dilution of anti-Flag antibody (Rabbit, Cell Signaling Technology) in antibody dilution buffer for

1.5 h at room temperature. After washing with PBS-TX twice for 5 min each, the samples were incubated with a 1:250 dilution of anti-Rabbit Alexa Fluor 594-conjugated secondary antibody (Invitrogen) and 1:6,000 dilution of GFP-dRNH1 at 1 mg/mL for 1.5 h at 37°C. After washing with PBS-TX for three times (5 min each), the samples were incubated with DAPI (Sigma) for 5 min to stain the DNA. The samples were mounted, imaged and quantified by following the aforementioned procedures.

Cell synchronization

The cell synchronization was conducted by following a previously described double thymidine block protocol with minor modifications (55). Pol κ-Flag HEK293T cells was plated at 20% confluency in a T75 culture flask containing 10 mL of DMEM supplemented with 10% fetal bovine serum (FBS) and 1% penicillin/streptomycin and incubated at 37°C overnight. Two hundred μL of 100 mM thymidine was added into the culture media until its final concentration reached 2 mM. The cells were incubated for 14 h and then washed with 10 mL pre-warmed PBS twice. The cells were subsequently incubated with 10 mL pre-warmed DMEM supplemented with 24 μM 2'-deoxycytidine for 9 h. Two hundred μL of 100 mM thymidine was added to the culture media again and the cells were incubated for 14 h. The cells were washed twice with 10 mL pre-warmed PBS and incubated in the pre-warmed DMEM supplemented with 24 μM 2'-deoxycytidine. The cells were then harvested at different time points for analysis of cell cycle by DNA staining using propidium iodide (PI, Sigma). The harvested cells were washed twice with 10 mL cold PBS and resuspended with the 0.1 mL PBS. The cells were fixed by adding 1 mL cold 80% ethanol dropwise with continuous vortexing and incubating on ice for 15 min. The cells were subsequently collected by centrifugation and stained with 500 μL PI/RNase A solution for 30 min. The samples were then analyzed with flow cytometry on Agilent NovoCyte Quanteon Flow Cytometer.

Results

To explore previously unrecognized functions of Pol κ, we employed an ascorbate peroxidase (APEX)-based proximity labeling method for assessing the interacting proteome of Pol κ in living cells (56, 57). To this end, we fused APEX2 to the C-terminus of Pol κ to facilitate the biotinylation of proximity proteins of Pol κ, where APEX2 linked with a nuclear localization signal (NLS) was chosen as a control. After transfection with plasmids for ectopic expression of Pol κ-APEX and NLS-APEX, HEK293T cells were treated with biotin-phenol (BP), followed by a 1.0-min exposure with H_2O_2 to catalyze the production of biotin-phenoxyl radicals, which react with electron-rich amino acids, e.g., tyrosine and cysteine, in nearby proteins (36). Immunoblotting revealed comparable expression levels of Pol κ-APEX and NLS-APEX, and similar levels of biotinylation of proteins in HEK293T cells elicited by the two fusion proteins (Figure 3-1). To minimize false-positive discovery of biotinylated proteins and to identify the domains in proximal proteins that interact with Pol κ, we employed an anti-biotin antibody to enrich the biotin-modified peptides (58), and subjected them to LC-MS/MS analysis (Figure 3-18a).

After database search, we detected over 2400 biotin-modified peptides from 1257 proteins from proximity labeling using Pol κ-APEX and NLS-APEX, where biotinylation occurs on tyrosine residues in most of these peptides (Figure 3-2a). We also detected biotinylation of other amino acid residues – including histidine, cysteine, and tryptophan – in proteins, which is in agreement with the previous report that the biotin-phenoxyl radical is capable of covalently conjugating with electron-rich amino acids (36, 59, 60). Gene Ontology analyses revealed that these proteins are mainly located in the nucleoplasm (Figure 3-2b). In addition, neither Pol κ-APEX nor NLS-APEX labeling led to detection of biotinylated peptides of GAPDH; however, both resulted in the detection of a comparable level of the biotinylated peptides at the same sites in lamin-B1 and

lamin-B2 (Figure 3-3), supporting the nuclear localization of Pol κ-APEX and NLS-APEX. Furthermore, XRCC1, a DNA repair protein known to interact with Pol κ (29), was exclusively detected in the Pol κ-APEX labeling sample (Figure 3-4), thereby validating the method.

A comparison of the LC-MS/MS data acquired from Pol κ-APEX and NLS-APEX labeling resulted in the identification of 280 proximity proteins of Pol κ (417 peptides, Figure 3-18b, Table S1), including many functionally related proteins of the polymerase, e.g., POLD1, histone H2AX, PARP1, and MCM6, etc. Interestingly, we also detected several RNA processing-related proteins, e.g., POLR2A, PAF1, TAF15, and RNase H2 (Table S1), suggesting that Pol κ may assume functions in RNA metabolism. Moreover, several helicases are selectively enriched in the Pol κ-APEX labeling samples compared with the control, including DDX1, DDX23, and DDX42 (Table S1), which is in line with the role of Pol κ in bypassing non-B form DNA (61).

The APEX labeling method involves ectopic overexpression of Pol κ, which may introduce artifacts in identifying its interaction proteins. Additionally, APEX method relies on biotinylation of amino acids located near the target protein of interest. Since only several electron-rich amino acids can be biotinylated, it is possible that some interaction partners of Pol κ, e.g., PCNA (a known interaction partner of Pol κ) (62, 63), may escape the labeling and subsequent LC-MS/MS detection. To overcome these limitations and to profile the interactome of endogenous Pol κ, we employed CRISPR-Cas9 to introduce three tandem repeats of Flag epitope tag to the C-terminus of endogenous Pol κ protein in HEK293T cells (Pol κ-Flag cells, Figure 3-5). Following enrichment using anti-Flag M2 beads and LC-MS/MS analysis, we identified 155 proteins with preferential binding toward Pol κ (Figure 3-18c, Table S1), including PCNA. In this vein, it is of note that we failed to detect XRCC1, a known binding partner of Pol κ (29), in the affinity pull-down mixture. This could be attributed to a limitation of the affinity pull-down technique, i.e., its inability to capture weak and/or transient interactions. Among these identified proteins, 36

overlap with the interacting proteins of Pol κ identified from the APEX proximity labeling experiment, including DDX23 (Figure 3-18d, Table S1).

We decided to conduct additional studies on DDX23 because our quantitative proteomic data unveiled a >30-fold enrichment of DDX23 in the Pol κ pull-down samples relative to the control (Table S1, Figure 3-6). We first validated the interaction between Pol κ and DDX23 by immunoprecipitation followed by Western blot analyses (Figure 3-19a). In addition, treatment with benzonase did not perturb the interaction between Pol κ and DDX23; hence, the interaction is independent of DNA or RNA (64). By using anti-Flag antibody, reciprocal immunoprecipitation of endogenous DDX23 from CRISPR-engineered HEK293T cells, which carry 3 tandem repeats of Flag epitope tag being incorporated to C-terminus of endogenous DDX23 (Figure 3-7), further confirmed its interaction with Pol κ in cells (Figure 3-19b).

To explore which domain(s) of DDX23 is (are) involved in its interaction with Pol κ, we overexpressed several truncated forms of HA-tagged DDX23 to perform the immunoprecipitation and Western blot analyses (41). The results showed that deletion of the transcription coactivator domain of DDX23 abrogated the interaction between Pol κ and DDX23 (Figure 3-19c, Figure 3-8), suggesting that Pol κ interacts DDX23 through the latter's transcription coactivator domain. Consistent with this notion, our APEX labeling results led to the identification of only one DDX23 peptide, which carries a biotinylated Y377 residing in the transcription coactivator domain (Figure 3-9, Figure 3-19d). To explore the domain of Pol κ involved in interaction with DDX23, we generated the EGFP-tagged wild-type Pol κ, along with its ubiquitin-binding domain (UBZ) mutant and catalytically dead (CD) mutant (42). Our results from co-IP experiments showed that Pol κ's UBZ domain, but not its catalytic activity, is required for its interaction with DDX23 (Figure 3-19e). Taken together, our two independent proteomic approaches and immunoprecipitation followed by Western blot analysis demonstrate an interaction between Pol κ

and DDX23, and the interaction entails the transcription coactivator domain of DDX23 and the UBZ domain of Pol κ.

Helicases assume functions in many steps of gene expression by unwinding double-stranded DNA/RNA and more complex nucleic acid structures, such as G quadruplexes and R-loops (65, 66). In particular, DDX23 was found to suppress R-loop formation and maintain genome stability (67). Moreover, DDX23 was previously identified as an R-loop-binding protein in two independent studies (68, 69). R-loops are nucleic acid structures formed by an RNA:DNA hybrid and a displaced single-stranded DNA (70). They arise when the nascent RNA molecule hybridizes with the template DNA strand during RNA polymerase II (RNAPII)-mediated transcription (71, 72). R-loops play crucial regulatory roles in a variety of cellular processes, including gene expression and DNA repair (73). Unscheduled R-loops, however, also compromise genomic stability and promote cancer development; hence, R-loops should be maintained at physiological levels (61, 74). Moreover, Pol κ could bypass unusual nucleic acid structures in cells (25). Thus, we hypothesized that Pol κ may function together with DDX23 in promoting R-loop resolution in cells.

To test this hypothesis, we first assessed the genome-wide occupancy of Pol κ using ChIP-Seq analysis with the aforementioned Pol κ-Flag cells. After peak calling with the control, we reproducibly identified 3991 Pol κ peaks from two biological replicates (Figures 3-10 and 3-20). Analysis of the Pol κ ChIP-Seq peaks showed that more than half of them (2009, 50.3%) reside in promoter regions (±1 kb from TSS, Figures 3-11 and 3-20a), with additional Pol κ peaks being mapped to gene body, terminal regions, and intergenic regions. In this vein, Pol κ's enrichment in promoter regions parallels the previous observations made for R-loops (51). Importantly, over 30% of Pol κ peaks overlapped with R-loop loci (Figure 3-20b). Moreover, the Pol κ ChIP-Seq signal was highly enriched and centered at R-loop sites, and vice versa (Figure 3-20c, d). Integrative

Genomics Viewer (IGV) plots also revealed a high degree of co-localization between Pol κ occupancy and R-loop loci in chromatin (Figure 3-20e). Together, the ChIP-Seq results revealed an enrichment of Pol κ at R-loop loci at the genome-wide scale, suggesting that Pol κ plays a role in R-loop processing. In this vein, Pol κ also exhibits a high degree of co-localization with RNAPII in chromatin (Figure 3-20e). Additionally, nine biotinylated peptides of RPB1 (RNAPII subunit A) were detected in the Pol κ-APEX proximity labeling samples (Figure 3-4), supporting a proximal localization of Pol κ with RNAPII in cells.

To confirm the ChIP-seq results, we conducted fluorescence microscopy analysis to examine the co-localization of R-loops with Pol κ and DDX23 in cells (Figure 3-12), where cellular R-loops were monitored using recombinant GFP-tagged catalytically dead RNase H1 (GFP-dRNH1) (52). Our results showed that DDX23 exhibits a strong co-localization with R-loops. Similar analysis also revealed a co-localization of Pol κ with R-loops (Pearson's correlation coefficients = 0.1-0.4), albeit to a lower extent than that of DDX23 (75). Furthermore, we observed that wild-type and catalytically inactive mutant of Pol κ exhibit substantially stronger co-localization with R-loops than its UBZ domain mutant, indicating that the UBZ domain of Pol κ contributes to its co-localization with R-loops.

Based on the above ChIP-Seq and proteomics data, we reasoned that Pol κ may promote R-loop resolution by recruiting DDX23. To test this, we conducted ChIP-qPCR analysis to examine whether the occupancy of DDX23 at R-loop loci is modulated by Pol κ with the use of Pol κ knockout cells and parental HEK293T cells (qPCR primers are listed in Table 3-1) (30). Based on the Pol κ ChIP-Seq and R-ChIP-Seq results (51), we selected three genes displaying enrichment of R-loops and Pol κ occupancy in their promoters, along with a negative control, i.e., *SNRPN* gene which does not exhibit enrichment of R-loops or Pol κ in its promoter (76, 77), for DDX23 ChIP-qPCR analysis using anti-Flag antibody (Figure 3-13 and 3-21). Our results showed

enrichment of DDX23 at R-loop loci in the promoters of *JUN*, *LENG8*, and *PMS2* genes, but not *SNRPN* gene (Figure 3-21d). We also employed CRISPR-Cas9 to insert three tandem repeats of the Flag tag to the C-terminus of endogenous DDX23 protein in Pol κ-deficient HEK293T cells (Figure 3-7). ChIP-qPCR analysis of with the use of these cells showed that loss of Pol κ led to diminished enrichment of DDX23 at these R-loop loci, underscoring a Pol κ-dependent recruitment of DDX23 to R-loop sites in chromatin. We also observed much more pronounced enrichment of Pol κ at R-loop loci in S phase cells than those out of the S-phase (Figures 3-14 and 3-15), suggesting a DNA replication-dependent recruitment of Pol κ to R-loop structures. To investigate whether DDX23 has a role in the recruitment of Pol κ to R-loop loci, we performed ChIP-qPCR analysis of Pol κ in HEK293T cells and the isogenic DDX23-deficient cells. The results showed that DDX23 is dispensable for the enrichment of Pol κ at R-loop loci (Figure 3-14).

To explore the roles of DDX23 and Pol κ in R-loop resolution, we performed R-ChIP-seq analysis in parental HEK293T (WT) and isogenic DDX23- and Pol κ-knockout (KO) cells (78). Gene annotation analysis showed that around 70% of R-loop peaks are located in promoter regions (Figures 3-16 and 3-22a), which is in keeping with R-loop distribution and validates our results (51). Genetic ablation of Pol κ and DDX23 led to elevations of the identified R-loop peak numbers from 1468 (in HEK293T cells) to 2451 and 5945 (Figure 3-22b), respectively, suggesting that Pol κ and DDX23 suppress the accumulation of R-loop structures at the genome-wide scale. In addition, the R-ChIP-seq signal is highly enriched and centered at Pol κ peaks (Figure 3-22c). Moreover, R-ChIP signal at Pol κ-binding sites exhibits higher intensities in Pol κ- and DDX23-deficient cells than parental HEK293T cells. These results substantiate our notion that Pol κ and DDX23 are involved in R-loop resolution.

To verify the R-ChIP-seq results, we monitored R-loop formation within the promoter regions of the aforementioned genes (*JUN, LENG8, PMS2*, and *SNRPN*) using R-ChIP-qPCR analysis (Figure 3-23). Cells stably expressing the catalytically inactive mutant of RNase H1 (D210N) showed a higher ChIP recovery rate in the promoter regions of *JUN, LENG8*, and *PMS2* genes, but not *SNRPN* gene, than the control cells stably expressing a mutant RNase H1 incompetent in R-loop binding (WKKD mutant) (51), suggesting R-loop formation in the promoter regions of JUN, LENG8, and PMS2 genes (Figure 3-23). Individual ablations of Pol κ and DDX23 resulted in increased R-loop formation in the promoters of *JUN, LENG8*, and *PMS2* genes, but not *SNRPN* gene (Figure 3-23).

To further examine the roles of DDX23 and Pol κ in R-loop resolution, we measured the global levels of R-loops in DDX23- and Pol κ-deficient cells by using fluorescence microscopy analysis with the above-mentioned GFP-dRNH1 (52). In line with the aforementioned results and a previous report (67), *DDX23* depletion led to markedly augmented accumulation of R-loops relative to the control (Figures 3-24a, b). Additionally, genetic ablation of Pol κ gave rise to elevated R-loop accumulation, albeit to a lesser degree than that of DDX23, suggesting the involvement of protein(s) other than Pol κ in aiding the recruitment of DDX23 to R-loop sites. Notably, the accumulation of R-loops in DDX23- and Pol κ-depleted cells was attenuated upon treatment with RNase H (Figure 3-24a, b). Together, these results suggest that Pol κ recruits DDX23 to promote R-loop resolution in cells.

We next asked whether Pol κ and DDX23's functions in R-loop resolution require their enzymatic activities. To this end, we complemented *POLK*$^{-/-}$ and *DDX23*$^{-/-}$ cells with wild-type Pol κ and DDX23, respectively, or their corresponding catalytically dead mutants (42, 54), and monitored R-loop levels in these cells. Our results revealed that both the polymerase activity of Pol κ and the helicase activity of DDX23 are required for R-loop resolution (Figure 3-24c-f). Furthermore,

consistent with the immunoprecipitation followed by Western blot results, the UBZ domain of Pol κ and the transcription coactivator domain of DDX23 contribute to their functions in R-loop resolution (Figure 3-24c-f). These findings parallel the above-mentioned results that Pol κ assumes elevated occupancies at R-loop-containing promoters in S-phase than non-S-phase cells, substantiating that Pol κ-DDX23-mediated R-loop resolution is likely accompanied with DNA replication.

Discussion

Recently, TLS polymerases were shown to have functions beyond their canonical roles in bypassing damaged DNA, where these polymerases emerged as new players in replication across structured DNA in vitro and in vivo (79, 80). Depletion of Pol κ increases cellular sensitivity toward telomestatin – a G4-stabilizing ligand – and confers genome instability (27). Pol κ was also shown to promote microsatellite stability and function on DNA templates containing sequences capable of adopting secondary structures within common fragile sites (81, 82). Collectively, these studies strongly suggest a role of Pol κ in replicating non-B form DNA.

By employing two independent proteomic strategies, relying on proximity labeling and affinity pull-down in conjunction with LC-MS/MS analysis, we explored comprehensively the interaction proteome of Pol κ. Our proteomic data revealed DDX23 as an interaction partner of Pol κ. We also validated the interaction between Pol κ and DDX23 by immunoprecipitation followed by Western blot analysis, and we found that the transcription coactivator domain of DDX23 and UBZ domain of Pol κ are indispensable for this interaction (Figure 3-19).

We also conducted the first ChIP-Seq analysis of endogenous Pol κ in HEK293T cells, and our results revealed the enrichment of the polymerase at R-loop structure loci. Moreover, our DDX23-ChIP-qPCR results showed diminished enrichment of DDX23 at R-loop loci upon

genetic ablation of Pol κ. We also validated the co-localizations of Pol κ and DDX23 with R-loops by using fluorescence imaging analysis. Moreover, R-ChIP-seq and fluorescence imaging results unveiled increased accumulation of R-loop structures after genetic depletion of Pol κ or DDX23. Based on these results, we propose a model where R-loop structures in template DNA stall replicative polymerases, which leads to the recruitment of Pol κ; the latter further recruits DDX23 to promote R-loop resolution (Figure 3-25).

Our proteomic data also provided an important resource for exploring other mechanisms that may modulate the functions of Pol κ through protein-protein interactions. For instance, we detected biotinylated peptides of RNase H2A and RNase H2B exclusively in the Pol κ-APEX labeling samples (Figure 3-4). We also observed enrichment of DNA topoisomerases 1 and 2 (Top1 and Top2) in the proximity proteome of Pol κ (Figure 3-4). These four proteins are crucial players in ribonucleotide excision repair (RER) (83-85); therefore, it will be important to examine whether Pol κ is also involved in RER.

Together, our multi-omics approach led to the discovery of a new function of Pol κ in R-loop resolution in human cells, documented its relevance at the genome-wide scale, and uncovered the involvement of DDX23 in this process. Hence, our work expanded the non-canonical functions of this highly conserved TLS polymerase. It will be interesting to examine whether this function of Pol κ can be extended to DNA damage-induced R-loops, if other TLS polymerases can promote R-loop resolution, and whether Pol κ-DDX23 interaction also contributes to replicative bypass of other structured DNA, including G-quadruplexes.

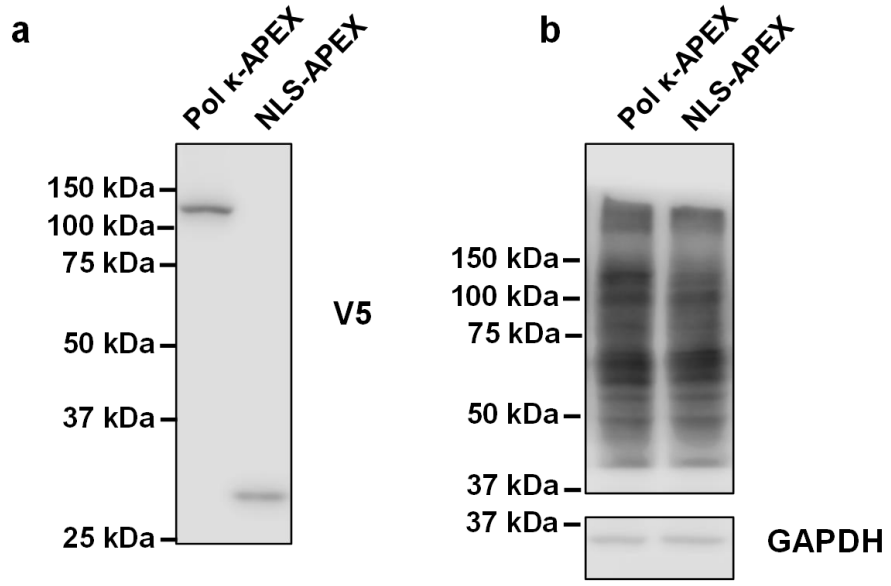

Figure 3-1. Western blot analyses of Pol κ-APEX and NLS-APEX expression (a) and labeling (b) in HEK293T cells.

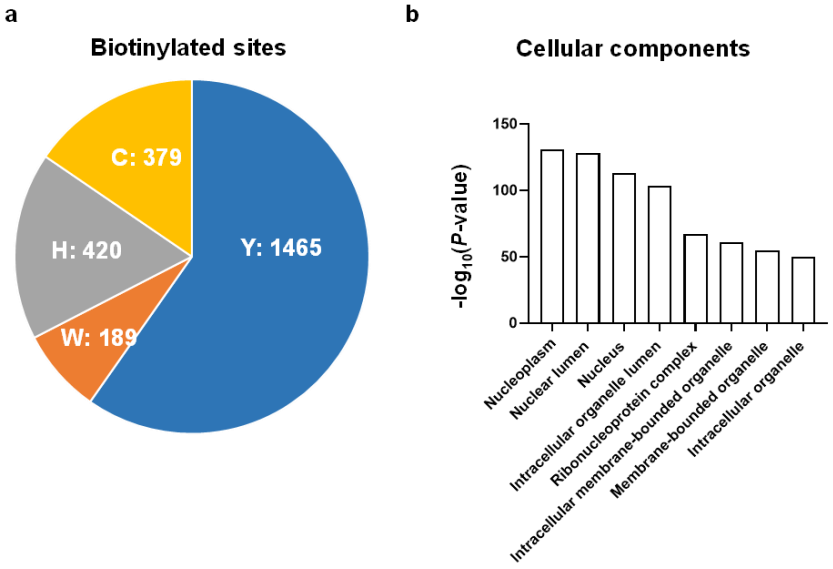

Figure 3-2. (a) The distribution of biotinylation sites at different amino acids residues, e.g., tyrosine (Y), histidine (H), cysteine (C), and tryptophan (W) identified from Pol κ-APEX labeling experiment. (b) GO analysis of proteins identified from APEX proteomics data.

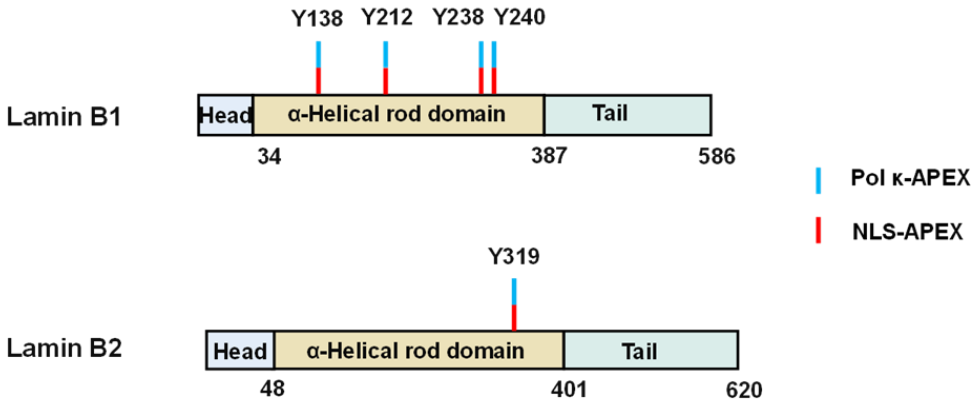

Figure 3-3. Biotinylated amino acids residues on lamin B1 and lamin B2 proteins detected with Pol κ-APEX and NLS-APEX.

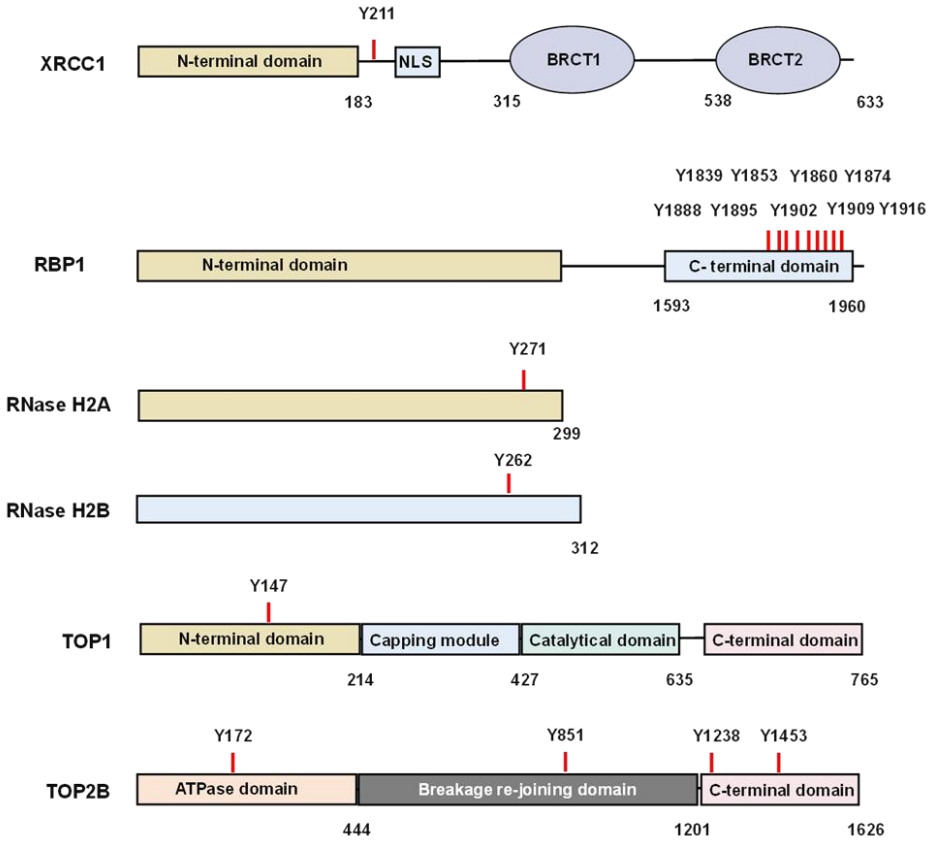

Figure 3-4. Biotinylated amino acids residues on selected proteins detected with Pol κ-APEX.

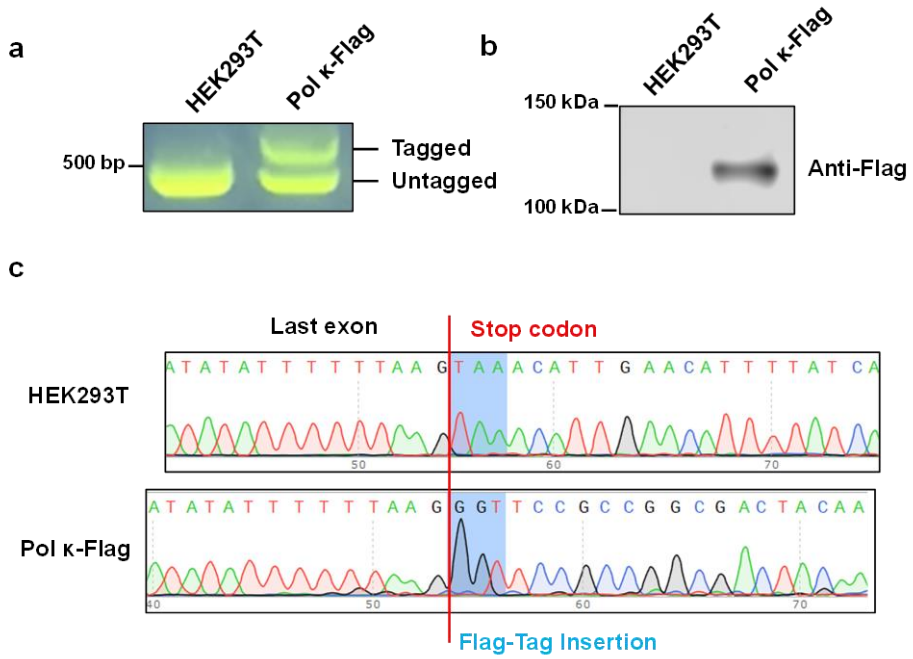

Figure 3-5. (a) Agarose gel electrophoresis for detecting targeted insertion of 3× Flag tag sequence to the C-terminus of Pol κ in HEK293T cells. PCR primers (Table 3-1) were designed to target outside the homology arms and yield a longer PCR product if the tag is inserted. The expected amplicon lengths of parental HEK293T cells and the knock-in clone are 424 and 601 bp, respectively. Shown also are the Western blot (b) and Sanger sequencing (c) results for confirming the successful knock-in (KI) of 3× Flag Tag to the C-terminus of endogenous Pol κ.

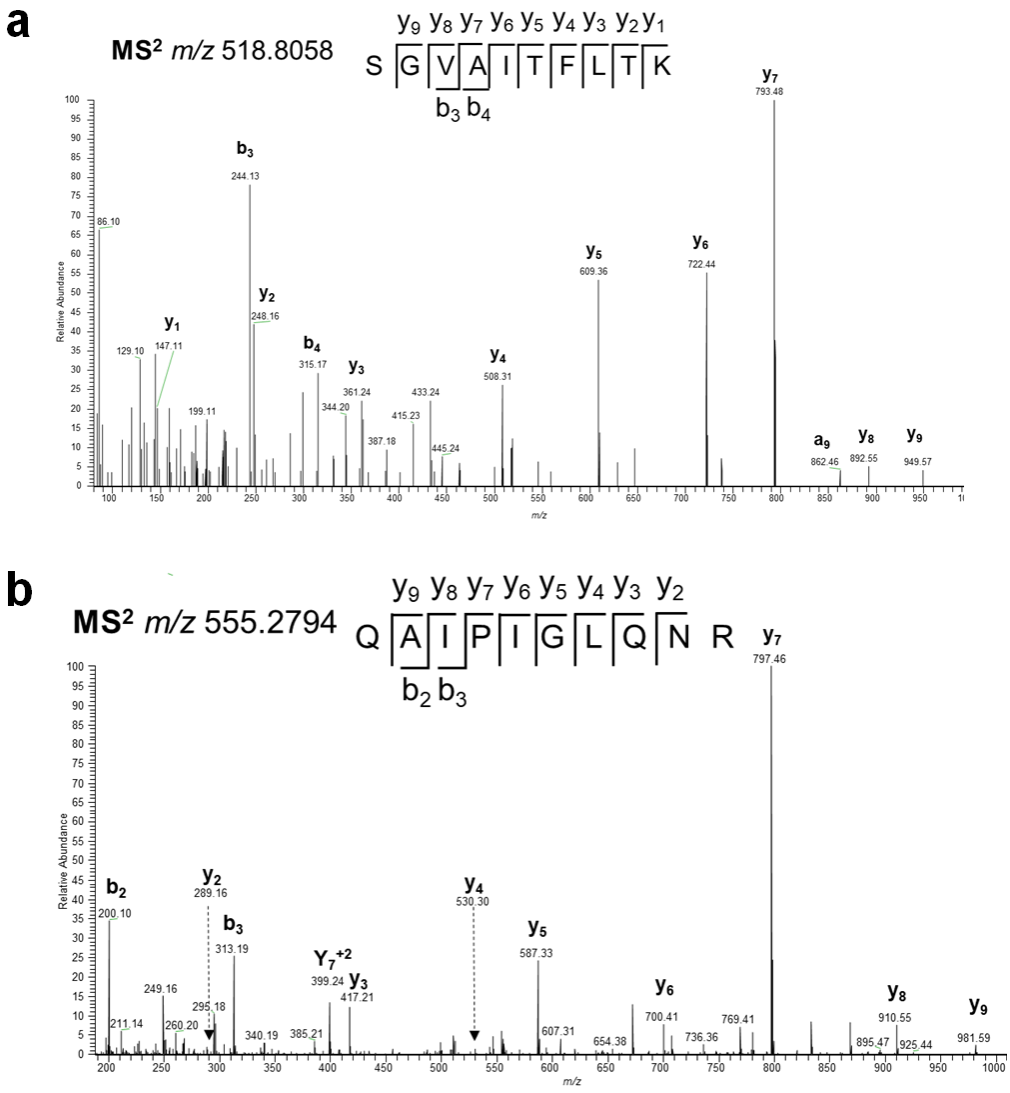

Figure 3-6. MS/MS for the $[M+2H]^{2+}$ ion of unique peptides of DDX23, e.g. (SGVAITFLTK) (a) and (QAIPIGLQNR) (b), identified from LC-MS/MS analysis of the tryptic digestion mixture of affinity pull-down sample using anti-Flag beads for Pol κ-Flag cells.

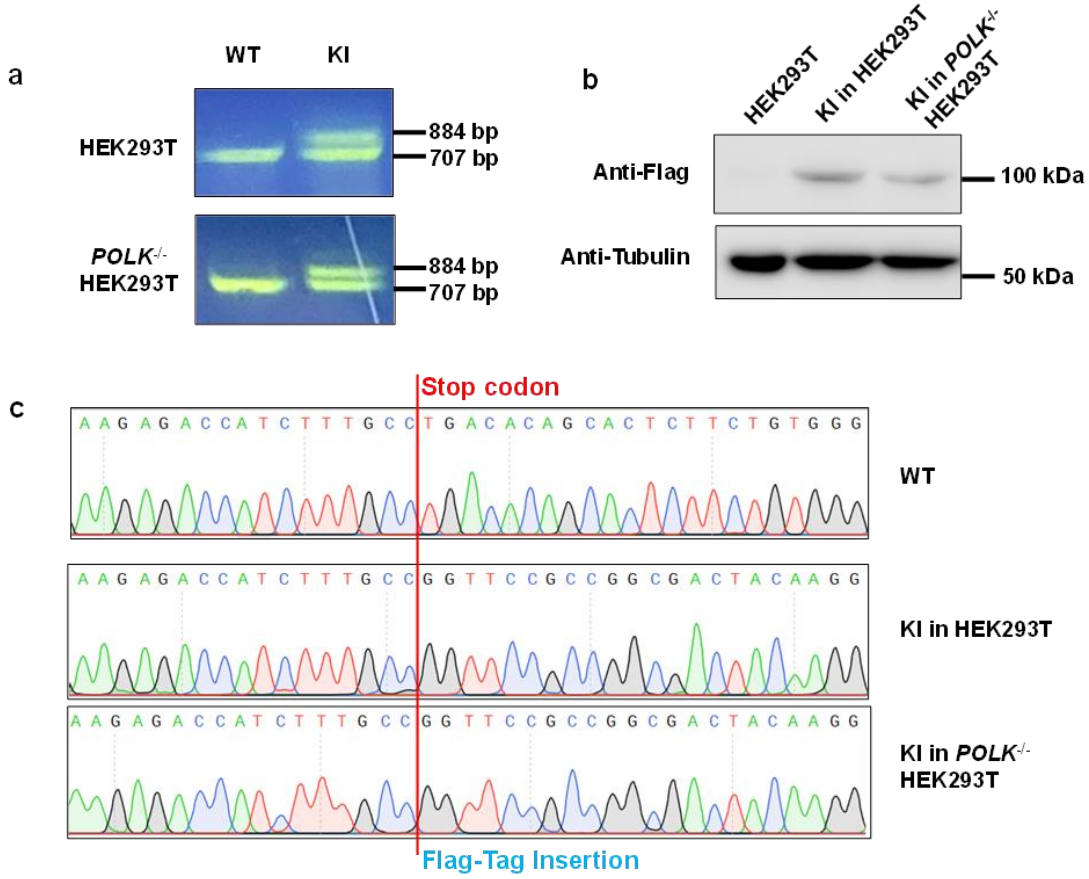

Figure 3-7. (a) Agarose gel electrophoresis for detecting targeted insertion of 3× Flag tag sequence to the C-terminus of DDX23 in HEK293T and Pol κ-deficient HEK293T cells. PCR primers (Table 3-1) were designed to target outside the homology arms and yield a longer PCR product if the tag is inserted. The expected amplicon lengths of parental HEK293T cells and the knock-in clone are 707 and 884 bp, respectively. Shown also are the Western blot (b) and Sanger sequencing (c) results for confirming the successful knockin (KI) of 3× Flag tag to the C-terminus of endogenous DDX23.

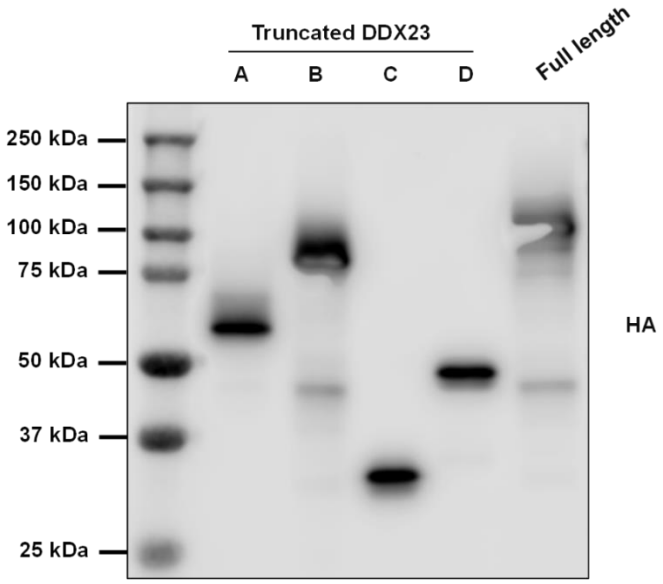

Figure 3-8. Western blot results showing the truncated DDX23 protein input blotted with anti-HA antibody.

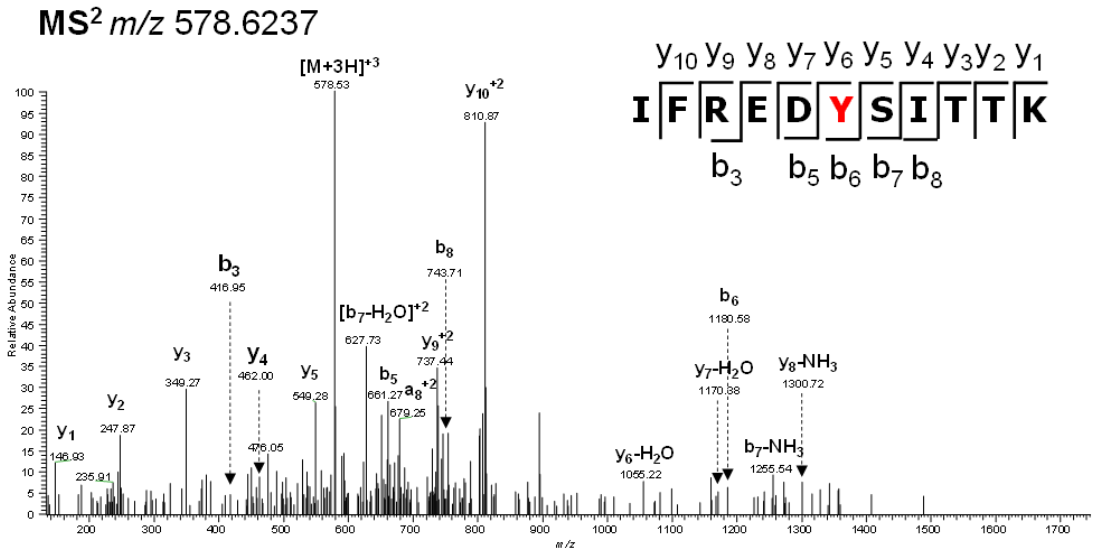

Figure 3-9. MS/MS for the $[M+3H]^{3+}$ of the biotinylated peptide (IFREDYSITTK) of DDX23 identified from the Pol κ APEX-labeling experiments, where the biotinylated tyrosine residue is highlighted in red.

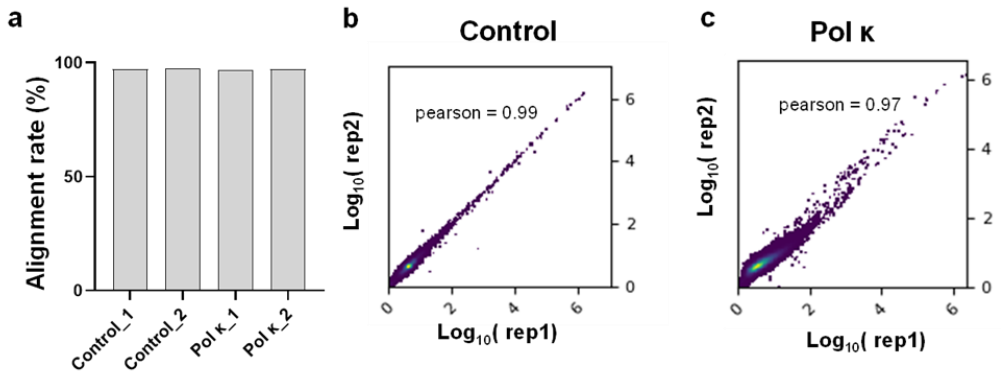

Figure 3-10. Alignment rates and correlation analyses of ChIP-Seq data. (a) The percentage of the sequencing reads that could be aligned with the hg19 human genome. Correlation analysis of sequencing data obtained from the two biological replicates of Control (b) and Pol κ (c).

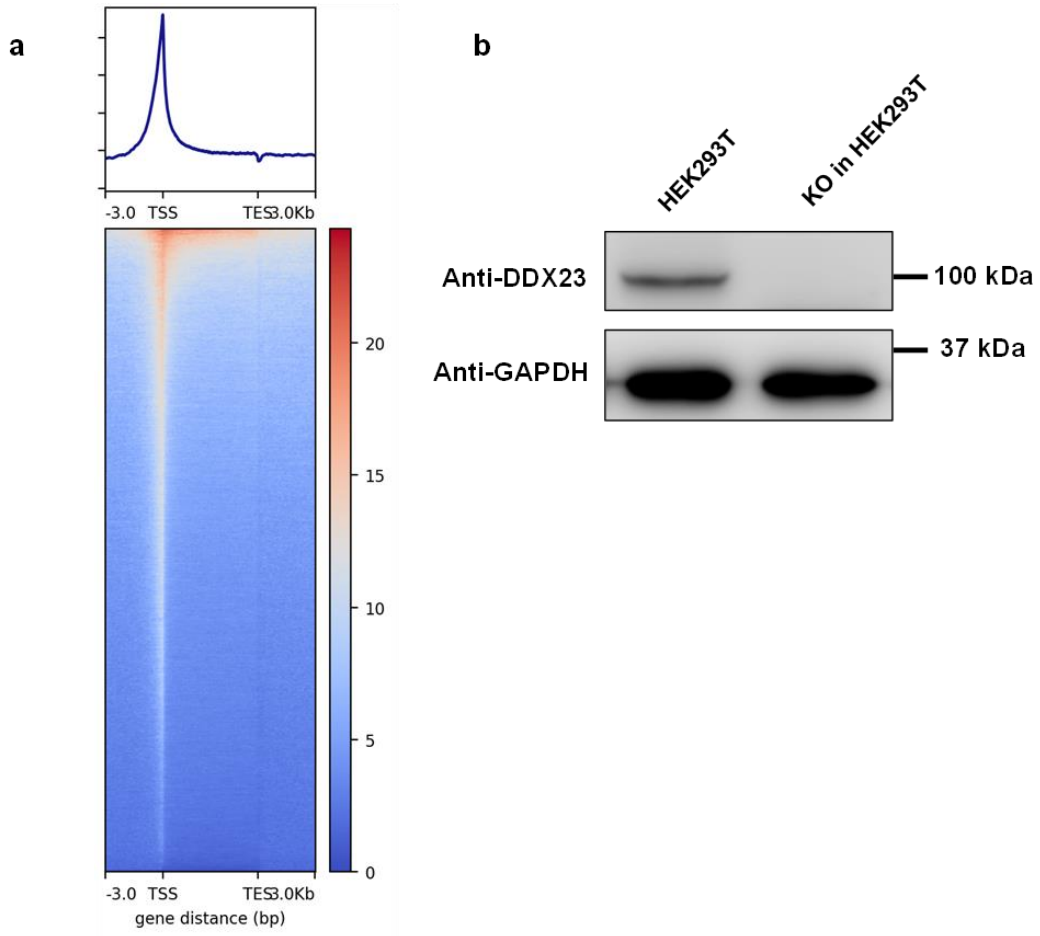

Figure 3-11. (a) Pol κ ChIP-Seq signal intensity meta profiles. (b) Western blot to confirm the successful knockout of DDX23 gene in HEK293T cells.

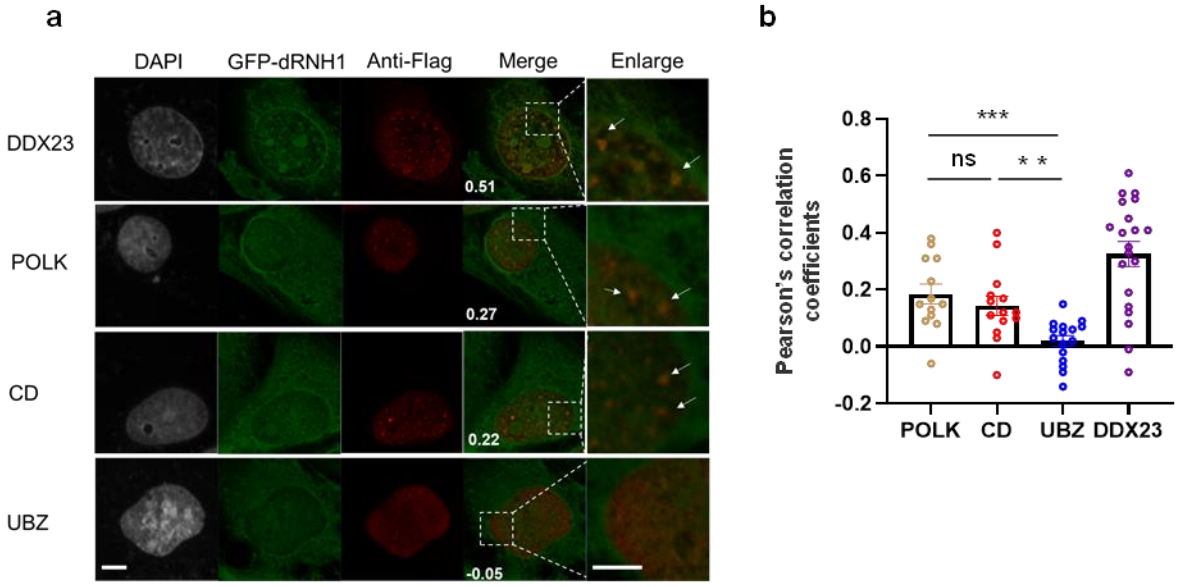

Figure 3-12. Fluorescence microscopy for analyzing the co-localization of R-loops with human Pol κ and DDX23. (a) Fluorescence microscopy analysis of R-loops in U2OS cells transfected with Flag-tagged DDX23, Pol κ, catalytically dead mutant of Pol κ (CD), and ubiquitin-binding domain mutant of Pol κ (UBZ). (b) Quantification of the Pearson's correlation coefficients between R-loops and different proteins indicated in (a). GFP-dRNH1, Flag and DAPI signals are shown in green, red, and blue, respectively. The data were obtained from at least 12 nuclei per condition. Statistical significance was determined using two-tailed, unpaired Student's t-test: ns, $p > 0.05$; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

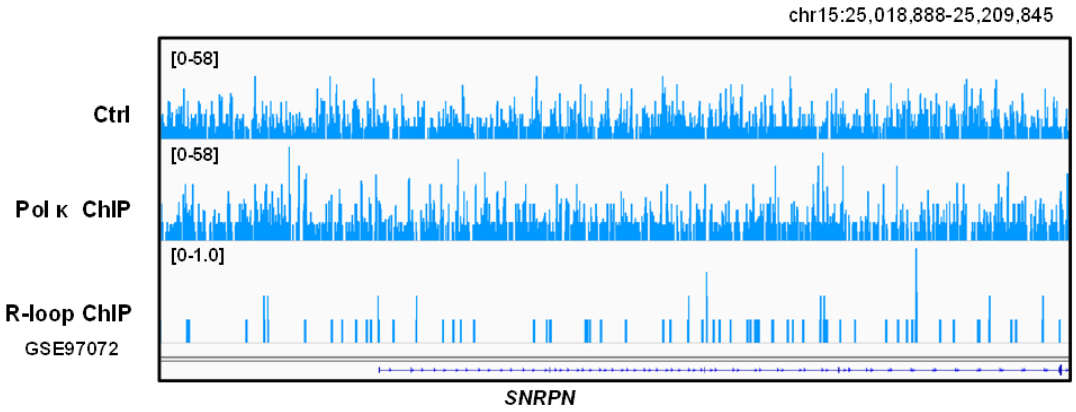

Figure 3-13. Pol κ and R-loop ChIP-Seq results for *SNRPN* gene in HEK293T cells.

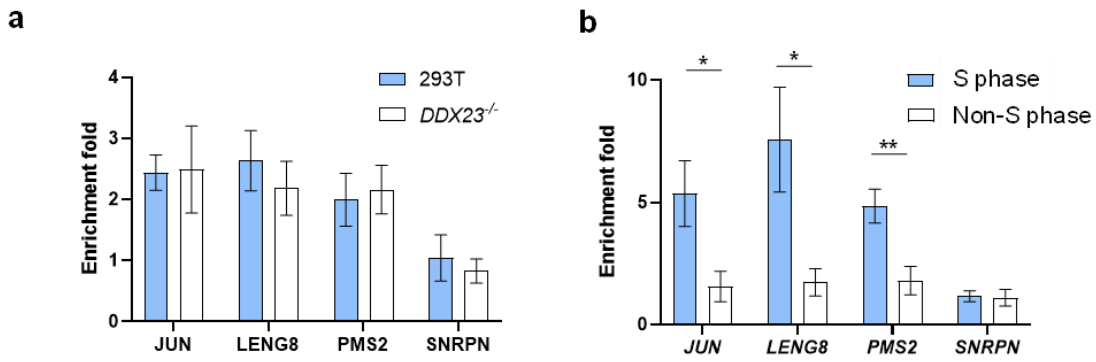

Figure 3-14. Pol κ-ChIP PCR analysis. (a) ChIP-qPCR analyses of Pol κ in the promoter regions of *JUN*, *LENG8*, *PMS2*, and *SNRPN* genes in HEK293T and *DDX23*$^{-/-}$ cells. (b) ChIP-qPCR analyses of Pol κ in the promoter regions of *JUN*, *LENG8*, *PMS2*, and *SNRPN* genes in HEK293T cells that were synchronized to the S phase and non-S phase, where the cells were collected at 1.5 and 9 h after release from double thymidine block, respectively (see Figure S15). The results were calculated as the enrichment fold by comparing to the control and presented as mean ± S.D. from three independent experiments. The *p* values were calculated by using two-tailed, unpaired Student's t-test: *, $p < 0.05$; **, $p < 0.01$.

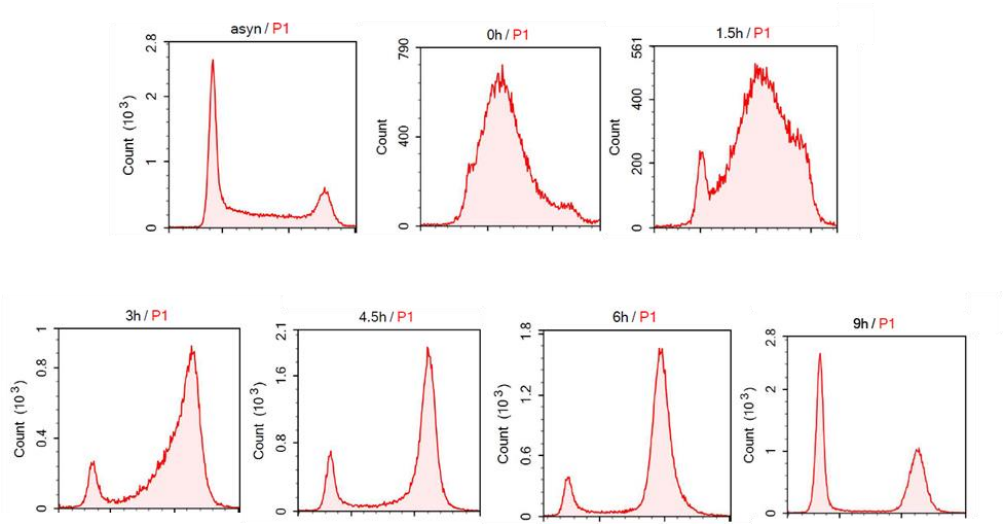

Figure 3-15. Cell cycle profiles at indicated time points (asynchronized, 0, 1.5, 3, 4.5, 6, 9 h) after release from double thymidine block.

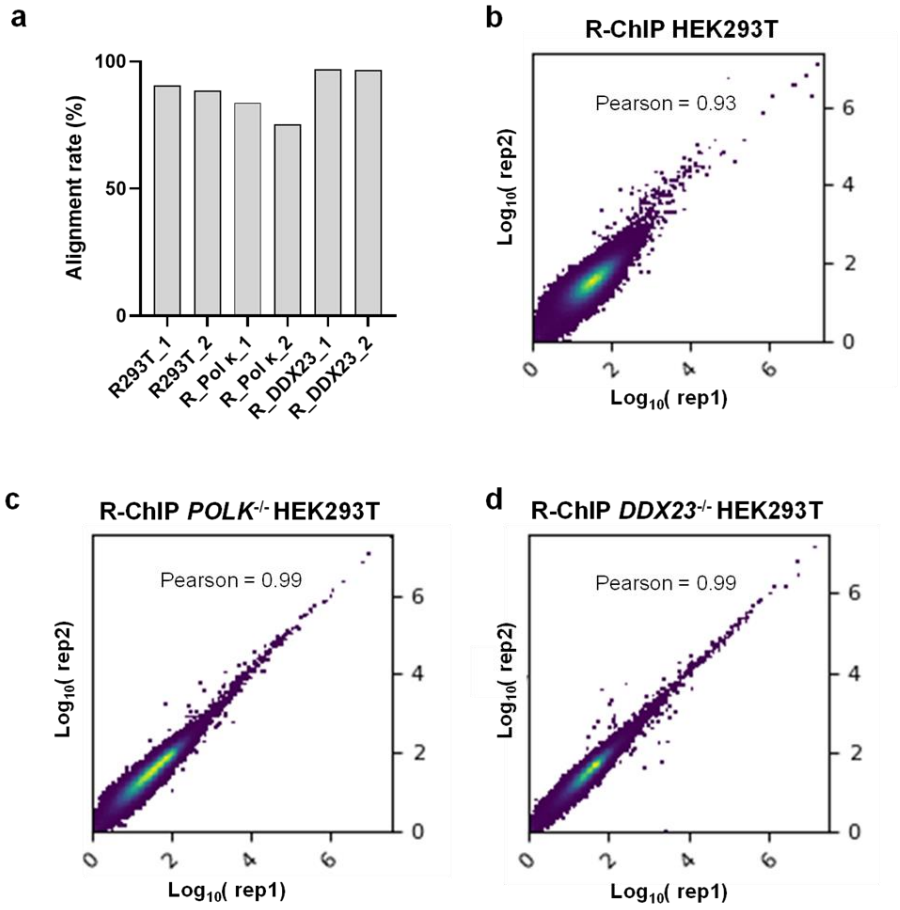

Figure 3-16. Alignment rates and correlation analyses of R-ChIP-Seq data. (a) The percentages of the sequencing reads that could be aligned with the hg19 human genome. Correlation analysis of sequencing data obtained from the two biological replicates of HEK293T (b), as well as the isogenic *POLK*$^{-/-}$ (c) and *DDX23*$^{-/-}$ (d) cells.

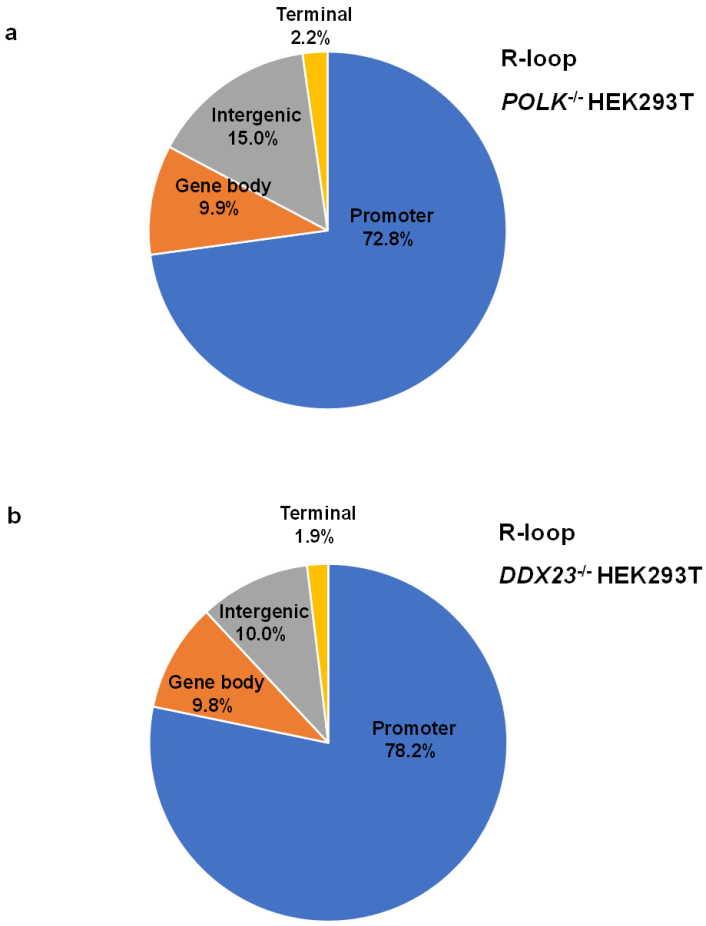

Figure 3-17. The genomic distribution of R-loop loci obtained from R-ChIP-Seq in *POLK*$^{-/-}$ and *DDX23*$^{-/-}$ HEK293T cells. Various genomic features are color coded according to the labels on the bottom.

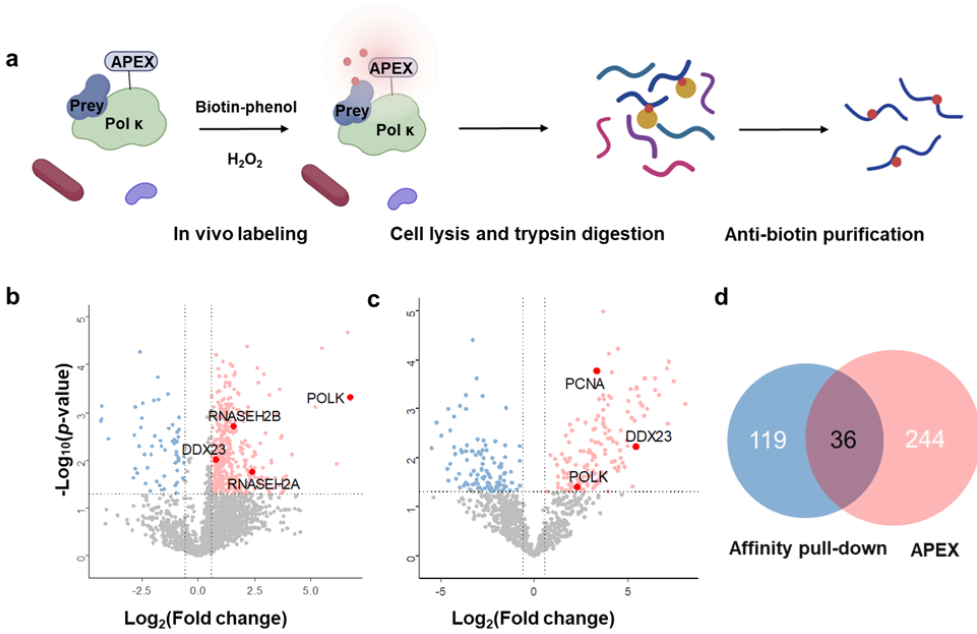

Figure 3-18. APEX labeling and affinity pull-down for revealing the Pol κ interactome. (a) Workflow for the enrichment and identification of biotinylated tryptic peptides induced by Pol κ-APEX. (b, c) A volcano plot displaying enriched proteins (highlighted in red) for Pol κ-APEX versus control (NLS-APEX) obtained from APEX labeling and LC-MS/MS analysis (b), or anti-Flag pull-down using CRISPR-engineered Pol κ-Flag cells versus parental HEK293T cells followed by LC-MS/MS analysis (c). Proteins with a > 1.5-fold signal over control and a *p* value < 0.05 are considered enriched. (d) A Venn diagram showing the candidate Pol κ-interaction proteins based on proteomic data obtained from affinity pull-down versus APEX-labeling.

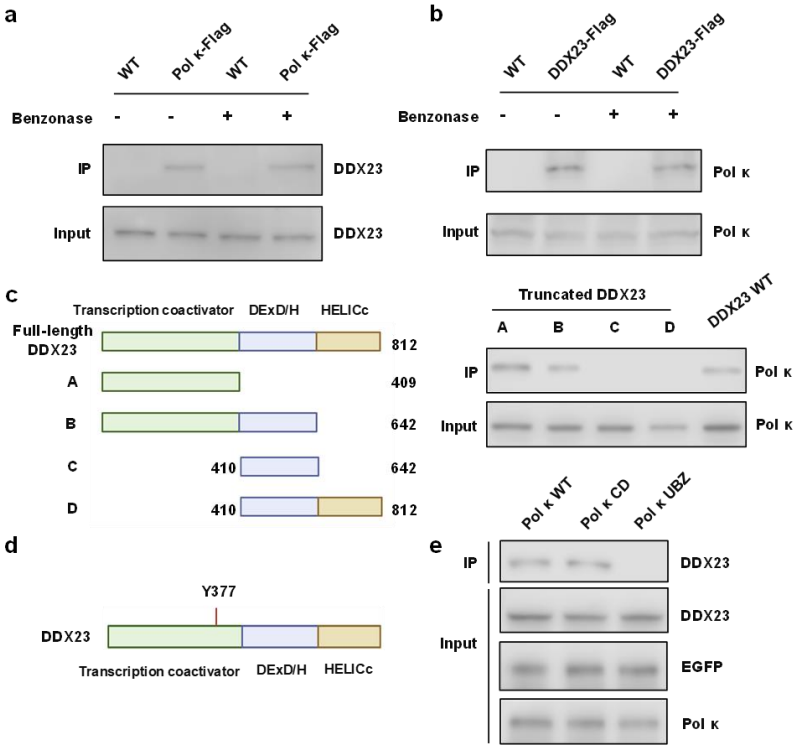

Figure 3-19. Immunoprecipitation followed by Western blot for assessing the interaction between DDX23 and Pol κ. (a) Western blot results of Flag pull-down samples using lysates of CRISPR-engineered Pol κ-Flag cells and parental HEK293T (WT) cells with or without benzonase treatment. (b) Western blot results for anti-Flag pull-down samples from lysates of CRISPR-engineered DDX23-Flag cells and parental HEK293T (WT) cells with or without benzonase treatment. (c) Western blot results for anti-HA pull-down samples from lysates of HEK293T cells transfected with full-length or truncated DDX23-HA. (d) Biotinylated amino acids residues of DDX23 detected in the Pol κ-APEX labeling. DExD/H-box and HELICc (helicase superfamily c-terminal) are highly conserved domains. (e) Western blot results showing anti-EGFP pull-down sample from lysates of HEK293T cells transfected with wild-type or mutant Pol κ-EGFP.

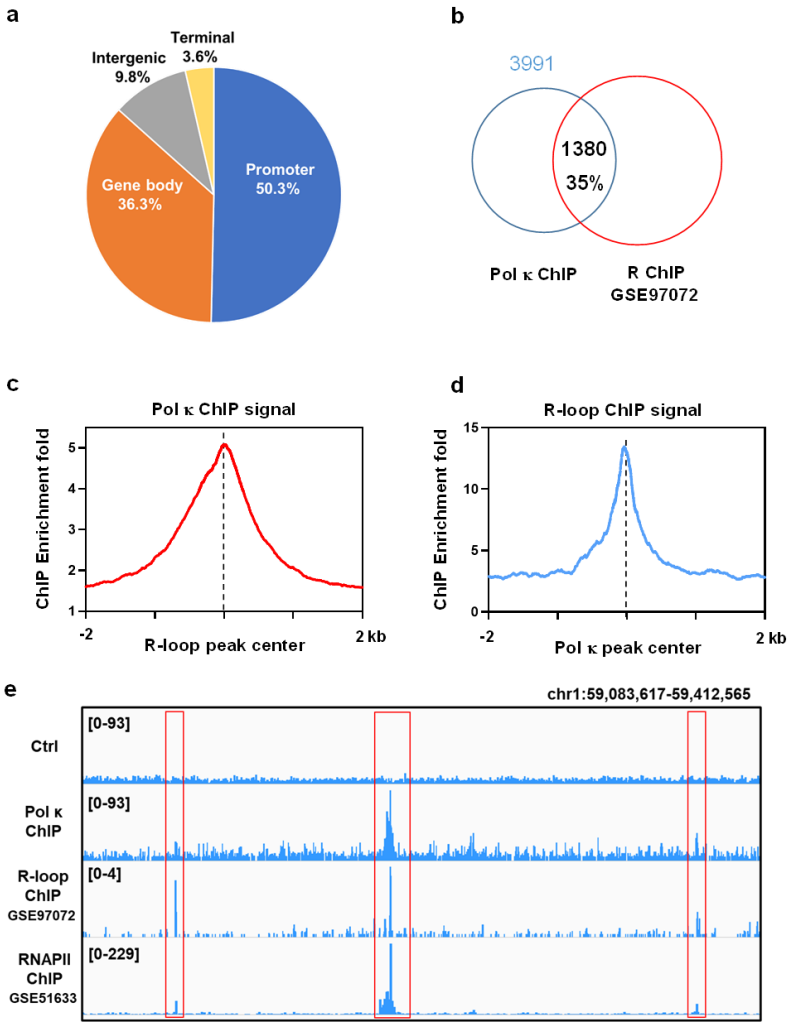

Figure 3-20. ChIP-seq revealed enrichment of Pol κ at R-loop structure loci in chromatin. (a) The genomic distribution of Pol κ-binding sites obtained from ChIP-Seq. (b) A Venn diagram depicting the overlap between Pol κ ChIP-seq and R-loop ChIP-seq peaks. (c) The average signal of Pol κ ChIP-seq against the center of the R-loop ChIP-seq peaks, and (d) vice versa. (e) A representative genomic region showing the co-localization among Pol κ occupancy, R-loop loci, and RNAPII occupancy.

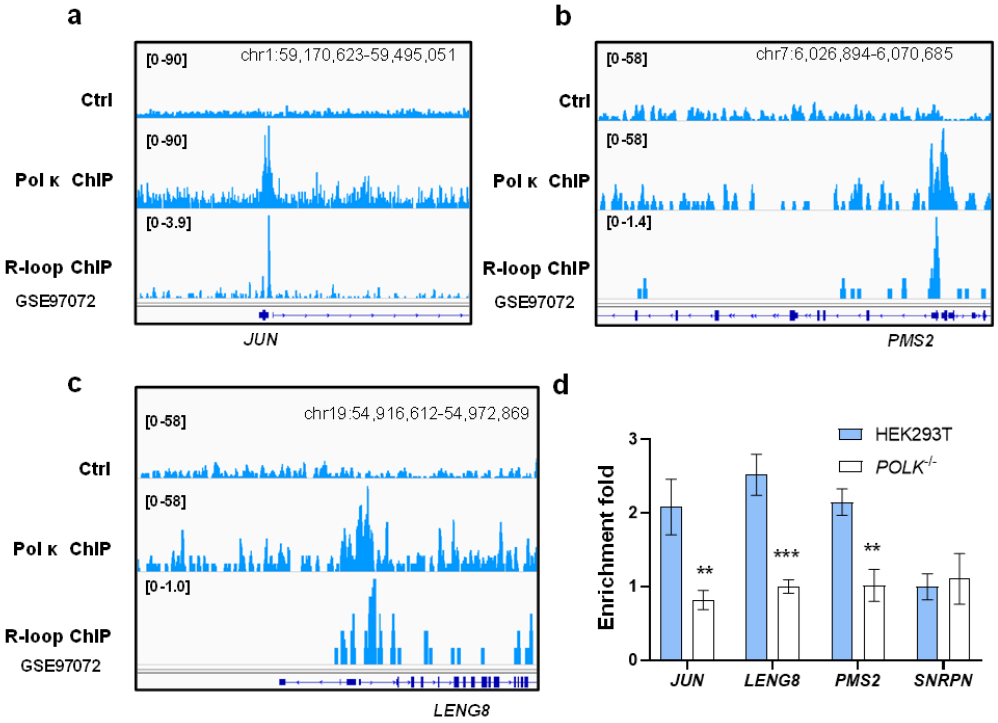

Figure 3-21. DDX23 is co-localized with Pol κ and R-loop structures. Representative IGV plots showing signal tracks for Pol κ and R-loop ChIP-seq results for *JUN* (a), *PMS2* (b), and *LENG8* (c) genes in HEK293T cells. (d) ChIP-qPCR analyses of DDX23 in the promoter regions of *JUN*, *LENG8*, *PMS2*, and *SNRPN* genes in HEK293T cells. Results were calculated as the enrichment fold by comparing to the control and represented as mean ± SD from three independent experiments. The *p* values were calculated by using two-tailed, unpaired Student's *t*-test: *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

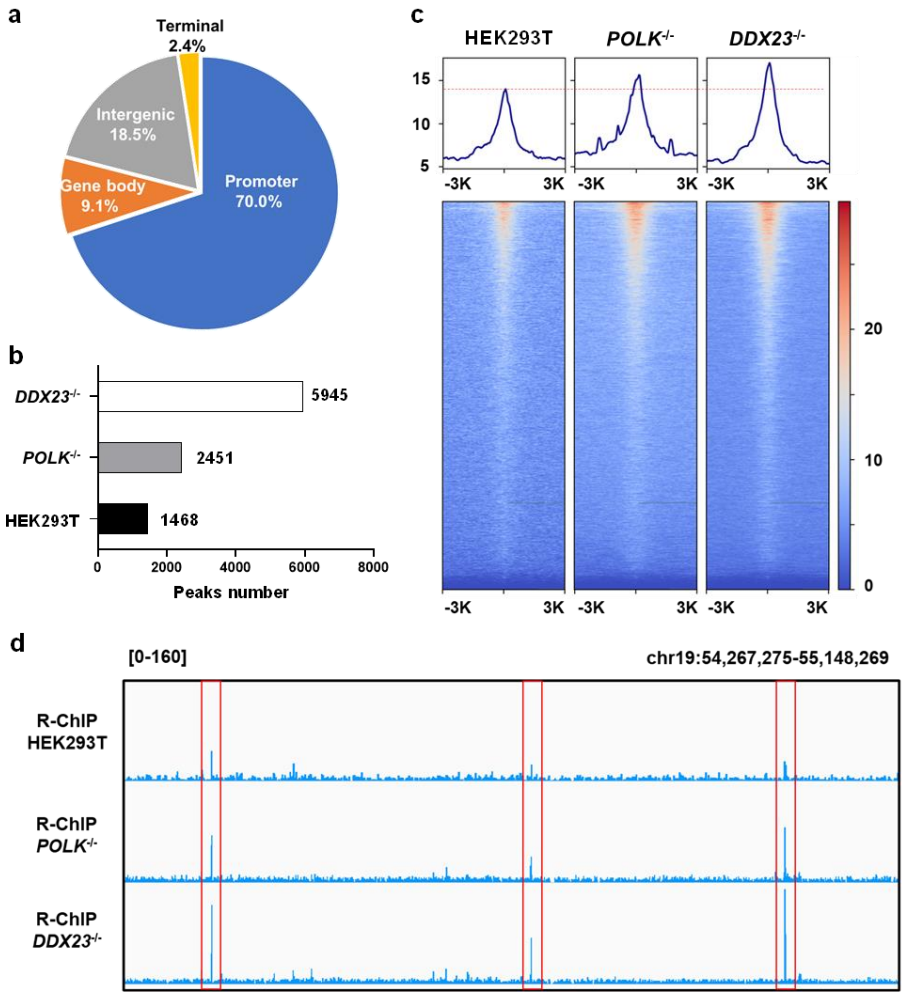

Figure 3-22. R-ChIP-seq revealed increased accumulation of R-loops in DDX23- and Pol κ-deficient HEK293T cells. (a) The genomic distribution of R-loop peaks mapped by R-ChIP-seq. (b) Peak numbers obtained from R-ChIP-seq experiments conducted for the three different cell lines. (c) The heatmaps of R-ChIP-seq signals obtained from three cell lines in regions within ± 3 kb from Pol κ peak centers.

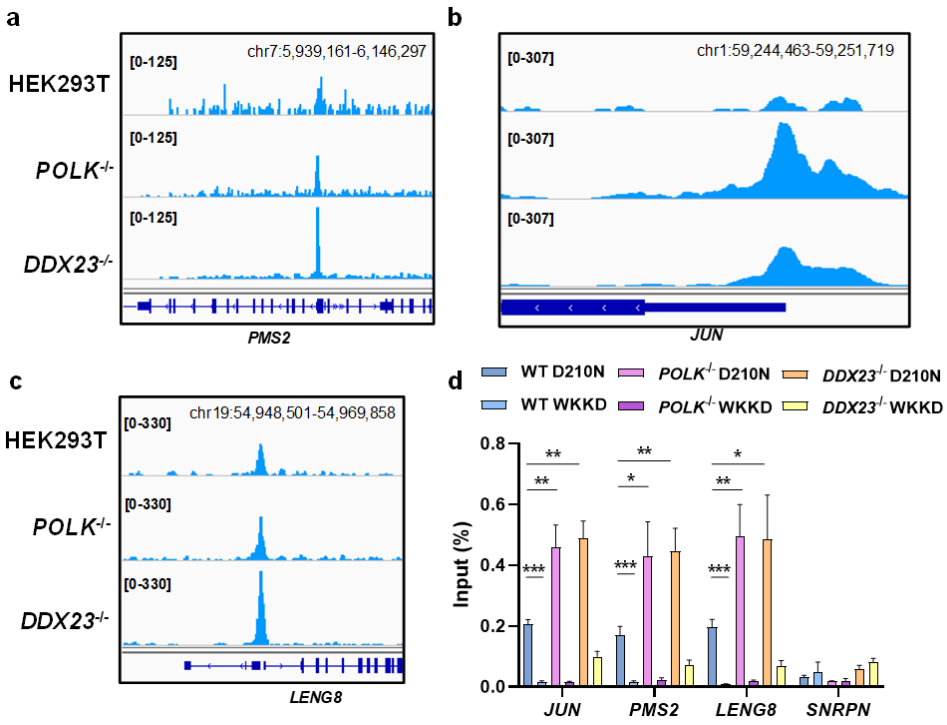

Figure 3-23. R-loop structures are accumulated in *POLK*$^{-/-}$ and *DDX23*$^{-/-}$ cells. Representative IGV plots showing signal tracks for R-loop ChIP-seq results for *PMS2* (a), *JUN* (b), and *LENG8* (c) genes in HEK293T and the isogenic *POLK*$^{-/-}$ and *DDX23*$^{-/-}$ cells. (d) R-ChIP-qPCR analyses of the promoter regions of *JUN*, *LENG8*, *PMS2*, and *SNRPN* genes in HEK293T, *POLK*$^{-/-}$, and *DDX23*$^{-/-}$ cells transfected with mutant (D210N and WKKD) RNase H1. The results were calculated as the percentage of input and presented as mean ± S.D. from three independent experiments. The *p* values were calculated by using two-tailed, unpaired Student's *t*-test: *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

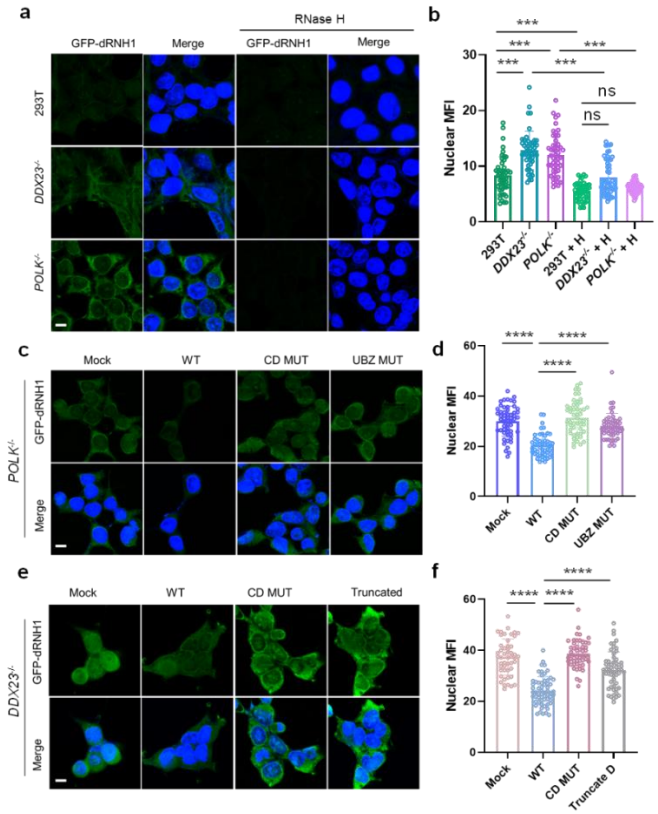

Figure 3-24. Fluorescence imaging analysis revealed increased accumulation of R-loops in the DDX23- and Pol κ-deficient HEK293T cells. (a) Fluorescence imaging analysis of R-loops in HEK293T, and the isogenic *DDX23*$^{-/-}$ and *POLK*$^{-/-}$cells. After fixation, coverslips were treated with or without RNase H, and then incubated with GFP-dRNH1. (b) Quantification of nuclear mean fluorescence intensities (MFI) for the conditions shown in (a). (c) Fluorescence imaging analysis of R-loops in Pol κ-deficient HEK293T cells complemented with vector (Mock), wild-type (WT), catalytically dead (CD) mutant, and ubiquitin-binding domain (UBZ) mutant of Pol κ. (d) Quantification of nuclear MFI for the conditions shown in (c). (e) Fluorescence imaging analysis of R-loops in DDX23-deficient HEK293T cells complemented with vector (Mock), wild-type (WT), catalytically dead (CD) mutant, and truncated D mutant of DDX23. (f) Quantification of nuclear MFI for the conditions shown in (e). GFP-dRNH1 and DAPI signals are shown in green and blue, respectively. Data were obtained from experiments conducted on two separate days, with 20 nuclei scored per condition per experiment (40 nuclei in total). Statistical significance was determined using two-tailed, unpaired Student's *t*-test: ns, $p > 0.05$; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; ****, $p < 0.0001$. Scale bars are 10 microns.

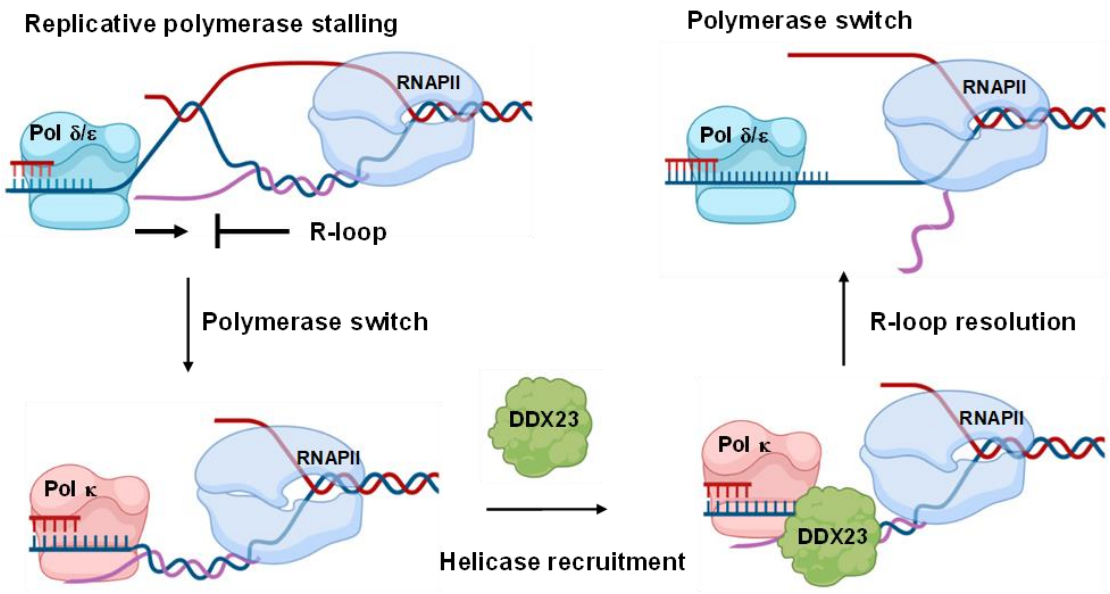

Figure 3-25. A proposed model for Pol κ's function in promoting R-loop resolution through DDX23. R-loops in template DNA stall replicative polymerases, which leads to switching of replicative polymerases with Pol κ; the latter further recruits DDX23 to promote the resolution of R-loop structures and facilitate their replicative bypass.

Primers used for ligation and RT-qPCR	
Sequence Names	Sequences
PMS2-F	5′- AGCTGAGAGCTCGAGGTGAG-3′
PMS2-R	5′- GAGATCGCTGCAACACTGAG-3′
JUN-F	5′- GGGTGACATCATGGGCTATT-3′
JUN-R	5′-TCGGACTATACTGCCGACCT-3′
LENG8-F	5′-CGCACTTACGCATGAACATT-3′
LENG8-R	5′- AGACTCCGTCTCCGAGAACA-3′
SNRPN-F	5′- GCTTTCCTTTCATTGGTTCAC-3′
SNRPN-R	5′- TTTCAAGGTACAGCTGGGAATA-3′
pRK7-POLK-F	5′- AAAAAGTCGACATGGATAGCAACAAAGGAG-3′
pRK7-POLK-R	5′- AAAAAGGATCCCTTAAAAAATATATCAAGGGTATG-3′
pRK7-DDX23-F	5′- AAAAATCTAGAATGGCAGGAGAGCTGGCTG-3′
pRK7-DDX23-R	5′- AAAAAGGATCCGGCAAAGATGGTCTCTTCCCGG-3′
DDX23-K441N-F	5′- GCTGAGACTGGCAGTGGCAACACAGCAGCC-3′
DDX23-K441N-R	5′- CTGCCAGTCTCAGCCACACCAATGATGTCACGA-3′
DDX23-A435V-F	5′- GTTGAGACTGGCAGTGGCAACACAGCAGCC-3′
DDX23-A435V-R	5′- CTGCCAGTCTCAACCACACCAATGATGTCACGA-3′
Sequences used for CRISPR/Cas9 editing	
POLK-sg-top	5′- CACCGTTACTTAAAAAATATATCAA -3′
POLK-sg-bottom	5′- AAACTTGATATATTTTTTAAGTAAC -3′
POLK-PCR-F	5′- AGGCCAGGATTGATGACAAAGT -3′
POLK -PCR-R	5′- TGTGTGAATCTTTGAGAGCCC -3′
DDX23-KO-sg-top	5′- CACCGCACGGACTCCTGACAGAGAG -3′
DDX23-KO-sg-bottom	5′- AAACCTCTCTGTCAGGAGTCCGTGC -3′
DDX23-KI-sg-top	5′- CACCGCTGACACAGCACTCTTCCTG -3′
DDX23-KI-sg-bottom	5′- AAACCAGGAAGAGTGCTGTGTCAGC -3′
DDX23-PCR-F	5′- TCAGTTCCAAGTCTGGGATTGC -3′
DDX23-PCR-R	5′- CTGGGCTAGGGAGTTGGATTG -3′

POLK-donor	5′- CCTCTAAAGGTAGCTCAAGTGGAGTACAGAAGGCTGTAACAAGAACA AAAAGGTATGGCTAATTTGAGCTTTAATAAAGCTGGCTACAATATGAA AATTCAATTATTTTATAAATTATTAAGTAATCAATATTAAAAATGAATTA TTTTCTAGAATCTCTTCCTGCATTTTAATTGTGTGTTCCTTTTCATTCT AGGCCAGGATTGATGACAAAGTACTCAACATCAAAGAAAATAAAACC AAACAATCCCAAACATACTCTTGATATATTTTTTAAGGGTTCCGCCGG CGACTACAAGGACCACGACGGCGATTATAAGGATCACGACATCGAC TACAAAGACGACGATGACAAGGGCGCCAGCAGCGCCTGGTCCCACC CTCAGTTTGAGAAGGGCGGAGGCTCTGGCGGCGGAAGCGGAGGAT CTGCTTGGAGCCACCCCCAGTTCGAAAAGTAAACATTGAACATTTTAT CATTAATTTTTAATTGAAACTAGTTATTTTATAATCAATGAATTTGTTCT TTCTGATTTTAAGTTTGCAGATTTATTTAGTGAAGGCAAGTGCAATAA TCCTTCCTCAGATGATGTTTGCTTTTCTAAGATACATATACTGATTCTG TGTATCTTTTTTATAACCATGAGAATTTTACTTCCATTATACATCAATT GGAAATCAATCCTGTTAAGAGATAATTCTTAAAAGGGAAATTAGGAAT GGGATAAGAAGGTGATTTTTTTATTATTTTTATACTGAATATAAAAACA TTTGTAAGGGCTCTCAAAGATTCACACATGCCTATATTATCATAAGAA TTTTTCAGCACTTAACTACTTTGTTGGCATTGATCCTAGTGTCTTTAAA TACTTCATGAGC -3′
DDX23-donor	5′- AGATGGGCACCTGGAGAGGGCACTGTGGGCGCCTCAGGTTCAGGC AGCATGTGCTGTATTCCTGCCTTTGGAGATGGGGTGGGAAGGATGC CTCTAGAAGTGGAGGCATCAGACACGGCTCCCCAGGGCATCTGGGC TATGCTGAGTGACGTTCTTCCCATTCCCCAGATTACATCCACCGCATT GGCCGCACGGGACGAGCAGGCAAGAGTGGGGTGGCCATCACCTTC CTCACAAAAGAGGACTCTGCTGTGTTCTACGAGCTGAAGCAAGCTAT CCTGGAAAGCCCAGTGTCTTCCTGTCCCCCCGAACTAGCCAACCAC CCAGATGCCCAGCATAAGCCAGGCACCATCCTCACCAAGAAGCGCC GGGAAGAGACCATCTTTGCCGGTTCCGCCGGCGACTACAAGGACCA CGACGGCGATTATAAGGATCACGACATCGACTACAAAGACGACGAT GACAAGGGCGCCAGCAGCGCCTGGTCCCACCCTCAGTTTGAGAAGG GCGGAGGCTCTGGCGGCGGAAGCGGAGGATCTGCTTGGAGCCACC CCCAGTTCGAAAAGTGACACAGCACTCTTCCTGTGGGCTGAGGGCA TCTCCAAAGCTGCCTGATGCCTGTTTTTCAGAACCCTCACATCCCTCT TTCCAGGTCCTCACTCTTGGGATATGGGGGCTTAGGAAAACAATCCA ACTCCCTAGCCCAGACCCTCAGGTCAGGAGGCCTGCGTGTGGGGCT GCAAAAGGAGAGGACGACGCTGTCGGAGGCAGGGAGAGCAAATTA CCACAGCTTCTTGGCCCAGTTCTGCCCTTCTTTGCTTTGGGATTGCA CTGGGCCATCAGCTCATGCCAGGCTATGGGGGCAGCCAGTTGGCAT TGCTCCCCAGACTGAACAGAAACCTGGCCGCCGGATGGGACCTCCT TT-3′

Table 3-1. Oligonucleotides used in this study.

References

1. T. Lindahl, Instability and decay of the primary structure of DNA. Nature 362, 709-715 (1993).

2. E. C. Friedberg, G. C. Walker, W. Siede et al., DNA Repair and Mutagenesis (ASM Press, Washington, D.C., 2006).

3. S. Prakash, R. E. Johnson, L. Prakash, Eukaryotic translesion synthesis DNA polymerases: specificity of structure and function. Annu. Rev. Biochem. 74, 317-353 (2005).

4. A. Vaisman, R. Woodgate, Translesion DNA polymerases in eukaryotes: what makes them tick? Crit. Rev. Biochem. Mol. Biol. 52, 274-303 (2017).

5. W. Yang, R. Woodgate, What a difference a decade makes: insights into translesion DNA synthesis. Proc. Natl. Acad. Sci. U.S.A. 104, 15591-15598 (2007).

6. W. Yang, Y. Gao, Translesion and Repair DNA Polymerases: Diverse Structure and Mechanism. Annu. Rev. Biochem. 87, 239-261 (2018).

7. C. Guo, J. N. Kosarek-Stancel, T. S. Tang, E. C. Friedberg, Y-family DNA polymerases in mammalian cells. Cell Mol. Life Sci. 66, 2363-2381 (2009).

8. R. P. Fuchs, S. Fujii, J. Wagner, Properties and functions of Escherichia coli: Pol IV and Pol V. DNA Repair and Replication 69, 229-264 (2004).

9. H. Ohmori, E. C. Friedberg, R. P. Fuchs et al., The Y-family of DNA polymerases. Mol. Cell 8, 7-8 (2001).

10. B. Yuan, H. Cao, Y. Jiang, H. Hong, Y. Wang, Efficient and accurate bypass of N2-(1-carboxyethyl)-2'-deoxyguanosine by DinB DNA polymerase in vitro and in vivo. Proc. Natl. Acad. Sci. U.S.A. 105, 8679-8684 (2008).

11. J. Wu, H. Du, L. Li et al., The impact of minor-groove N2-alkyl-2'-deoxyguanosine lesions on DNA replication in human cells. ACS Chem. Biol. 14, 1708-1716 (2019).

12. J. Y. Choi, K. C. Angel, F. P. Guengerich, Translesion synthesis across bulky N2-alkyl guanine DNA adducts by human DNA polymerase k. J. Biol. Chem. 281, 21062-21072 (2006).

13. D. F. Jarosz, V. G. Godoy, J. C. Delaney, J. M. Essigmann, G. C. Walker, A single amino acid governs enhanced activity of DinB DNA polymerases on damaged templates. Nature 439, 225-228 (2006).

14. Y. Liu, Y. Yang, T. S. Tang et al., Variants of mouse DNA polymerase k reveal a mechanism of efficient and accurate translesion synthesis past a benzo[a]pyrene dG adduct. Proc. Natl. Acad. Sci. U.S.A. 111, 1789-1794 (2014).

15. S. Avkin, M. Goldsmith, S. Velasco-Miguel et al., Quantitative analysis of translesion DNA synthesis across a benzo[a]pyrene-guanine adduct in mammalian cells: the role of DNA polymerase kappa. J. Biol. Chem. 279, 53298-53305 (2004).

16. T. Ogi, Y. Shinkai, K. Tanaka, H. Ohmori, Polk protects mammalian cells against the lethal and mutagenic effects of benzo[a]pyrene. Proc. Natl. Acad. Sci. U.S.A. 99, 15548-15553 (2002).

17. B. Yuan, C. You, N. Andersen et al., The roles of DNA polymerases k and i in the error-free bypass of N2-carboxyalkyl-2'-deoxyguanosine lesions in mammalian cells. J. Biol. Chem. 286, 17503-17511 (2011).

18. Y. Zhang, F. Yuan, X. Wu et al., Error-free and error-prone lesion bypass by human DNA polymerase kappa in vitro. Nucleic. Acids. Res. 28, 4138-4146 (2000).

19. J. H. Yoon, G. Bhatia, S. Prakash, L. Prakash, Error-free replicative bypass of thymine glycol by the combined action of DNA polymerases kappa and zeta in human cells. Proc. Natl. Acad. Sci. U.S.A. 107, 14116-14121 (2010).

20. P. L. Fischhaber, V. L. Gerlach, W. J. Feaver et al., Human DNA polymerase kappa bypasses and extends beyond thymine glycols during translesion synthesis in vitro, preferentially incorporating correct nucleotides. J. Biol. Chem. 277, 37604-37611 (2002).

21. Y. Kanemaru, T. Suzuki, A. Sassa et al., DNA polymerase k protects human cells against MMC-induced genotoxicity through error-free translesion DNA synthesis. Genes Environ. 39, 6 (2017).

22. M. Enoiu, J. Jiricny, O. D. Scharer, Repair of cisplatin-induced DNA interstrand crosslinks by a replication-independent pathway involving transcription-coupled repair and translesion synthesis. Nucleic Acids Res. 40, 8953-8964 (2012).

23. T. V. Ho, A. Guainazzi, S. B. Derkunt, M. Enoiu, O. D. Scharer, Structure-dependent bypass of DNA interstrand crosslinks by translesion synthesis polymerases. Nucleic Acids Res. 39, 7455-7464 (2011).

24. V. Jha, H. Ling, Structural basis for human DNA polymerase k to bypass cisplatin intrastrand cross-link (Pt-GG) lesion as an efficient and accurate extender. J. Mol. Biol. 430, 1577-1589 (2018).

25. E. Walsh, X. Wang, M. Y. Lee, K. A. Eckert, Mechanism of replicative DNA polymerase d pausing and a potential role for DNA polymerase k in common fragile site replication. J. Mol. Biol. 425, 232-243 (2013).

26. K. L. Burr, S. Velasco-Miguel, V. S. Duvvuri et al., Elevated mutation rates in the germline of Polk mutant male mice. DNA Repair 5, 860-862 (2006).

27. R. Betous, L. Rey, G. Wang et al., Role of TLS DNA polymerases h and k in processing naturally occurring structured DNA in human cells. Mol. Carcinog. 48, 369-378 (2009).

28. Y. Kanemaru, T. Suzuki, N. Niimi et al., Catalytic and non-catalytic roles of DNA polymerase k in the protection of human cells against genotoxic stresses. Environ. Mol. Mutagen. 56, 650-662 (2015).

29. T. Ogi, S. Limsirichaikul, R. M. Overmeer et al., Three DNA polymerases, recruited by different mechanisms, carry out NER repair synthesis in human cells. Mol. Cell 37, 714-727 (2010).

30. J. Wu, L. Li, P. Wang et al., Translesion synthesis of O4-alkylthymidine lesions in human cells. Nucleic Acids Res. 44, 9256-9265 (2016).

31. F. Tang, J. Yuan, B. F. Yuan, Y. Wang, DNA-protein cross-linking sequencing for genome-wide mapping of thymidine glycol. J. Am. Chem. Soc. 144, 454-462 (2022).

32. F. A. Ran, P. D. Hsu, J. Wright et al., Genome engineering using the CRISPR-Cas9 system. Nat. Protoc. 8, 2281-2308 (2013).

33. F. Tang, Y. Wang, Z. Gao, S. Guo, Y. Wang, Polymerase eta recruits DHX9 helicase to promote replication across guanine quadruplex structures. J. Am. Chem. Soc. 144, 14016-14020 (2022).

34. L. B. Alexandrov, P. H. Jones, D. C. Wedge et al., Clock-like mutational processes in human somatic cells. Nat. Genet. 47, 1402-+ (2015).

35. P. Kaewsapsak, D. M. Shechner, W. Mallard, J. L. Rinn, A. Y. Ting, Live-cell mapping of organelle-associated RNAs via proximity biotinylation combined with protein-RNA crosslinking. eLife 6 (2017).

36. V. Hung, N. D. Udeshi, S. S. Lam et al., Spatially resolved proteomic mapping in living cells with the engineered peroxidase APEX2. Nat. Protoc. 11, 456-475 (2016).

37. T. F. Qi, X. Liu, F. Tang et al., Targeted quantitative profiling of epitranscriptomic reader, writer, and eraser proteins using stable isotope-labeled peptides. Anal. Chem. 94, 12559-12564 (2022).

38. S. Tyanova, T. Temu, J. Cox, The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nat. Protoc. 11, 2301-2319 (2016).

39. A. D. Shah, R. J. A. Goode, C. Huang, D. R. Powell, R. B. Schittenhelm, LFQ-analyst: an easy-to-use interactive web platform to analyze and visualize label-free proteomics data preprocessed with MaxQuant. J. Proteome Res. 19, 204-211 (2020).

40. Y. Perez-Riverol, J. Bai, C. Bandla et al., The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences. Nucleic Acids Res. 50, D543-D552 (2022).

41. J. Ruan, Y. Cao, T. Ling et al., DDX23, an evolutionary conserved dsRNA sensor, participates in innate antiviral responses by pairing with TRIF or MAVS. Front. Immunol. 10, 2202 (2019).

42. P. Tonzi, Y. Yin, C. W. T. Lee, E. Rothenberg, T. T. Huang, Translesion polymerase kappa-dependent DNA synthesis underlies replication fork recovery. eLife 7 (2018).

43. F. Tang, S. Liu, Q. Y. Li et al., Location analysis of 8-oxo-7,8-dihydroguanine in DNA by polymerase-mediated differential coding. Chem. Sci. 10, 4272-4281 (2019).

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

45. J. Feng, T. Liu, B. Qin, Y. Zhang, X. S. Liu, Identifying ChIP-seq enrichment using MACS. Nat. Protoc. 7, 1728-1740 (2012).

46. J. W. Nicol, G. A. Helt, S. G. Blanchard, Jr., A. Raja, A. E. Loraine, The Integrated Genome Browser: free software for distribution and exploration of genome-scale datasets. Bioinformatics 25, 2730-2731 (2009).

47. S. Heinz, C. Benner, N. Spann et al., Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 38, 576-589 (2010).

48. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-842 (2010).

49. F. Ramirez, D. P. Ryan, B. Gruning et al., deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Res. 44, W160-165 (2016).

50. M. Dalvai, J. Loehr, K. Jacquet et al., A Scalable Genome-Editing-Based Approach for Mapping Multiprotein Complexes in Human Cells. Cell Rep. 13, 621-633 (2015).

51. L. Chen, J. Y. Chen, X. Zhang et al., R-ChIP using inactive RNase H reveals dynamic coupling of R-loops with transcriptional pausing at gene promoters. Mol. Cell 68, 745-757 (2017).

52. M. P. Crossley, J. R. Brickner, C. Song et al., Catalytically inactive, purified RNase H1: A specific and sensitive probe for RNA-DNA hybrid imaging. J. Cell Biol. 220 (2021).

53. X. Dai, C. You, Y. Wang, The Functions of Serine 687 Phosphorylation of Human DNA Polymerase eta in UV Damage Tolerance. Mol. Cell. Proteomics 15, 1913-1920 (2016).

54. A. Pause, N. Sonenberg, Mutational analysis of a DEAD box RNA helicase: the mammalian translation initiation factor eIF-4A. EMBO J. 11, 2643-2654 (1992).

55. H. T. Ma, R. Y. Poon, Synchronization of HeLa cells. Methods Mol. Biol. 761, 151-161 (2011).

56. H. W. Rhee, P. Zou, N. D. Udeshi et al., Proteomic mapping of mitochondria in living cells via spatially restricted enzymatic tagging. Science 339, 1328-1331 (2013).

57. F. Tang, Y. Wang, Z. Gao, S. Guo, Y. Wang, Polymerase h recruits DHX9 helicase to promote replication across guanine quadruplex structures. J. Am. Chem. Soc. 144, 14016-14020 (2022).

58. N. D. Udeshi, K. Pedram, T. Svinkina et al., Antibodies to biotin enable large-scale detection of biotinylation sites on proteins. Nat. Methods 14, 1167-1170 (2017).

59. B. Bhaskar, C. E. Immoos, H. Shimizu et al., A novel heme and peroxide-dependent tryptophan-tyrosine cross-link in a mutant of cytochrome c peroxidase. J. Mol. Biol. 328, 157-166 (2003).

60. F. Amini, T. Kodadek, K. C. Brown, Protein affinity labeling mediated by genetically encoded peptide tags. Angew. Chem. Int. Ed. Engl. 41, 356-359 (2002).

61. M. J. Pillaire, R. Betous, J. S. Hoffmann, Role of DNA polymerase k in the maintenance of genomic stability. Mol. Cell Oncol. 1, e29902 (2014).

62. C. Lancey, M. Tehseen, S. Bakshi et al., Cryo-EM structure of human pol k bound to DNA and mono-ubiquitylated PCNA. Nat. Commun. 12, 6095 (2021).

63. M. Bienko, C. M. Green, N. Crosetto et al., Ubiquitin-binding domains in Y-family polymerases regulate translesion synthesis. Science 310, 1821-1824 (2005).

64. G. Baytek, O. Popp, P. Mertins, B. Tursun, Robust co-immunoprecipitation with mass spectrometry for Caenorhabditis elegans using solid-phase enhanced sample preparation. Biotechniques. 72, 175-184 (2022).

65. M. R. Singleton, M. S. Dillingham, D. B. Wigley, Structure and mechanism of helicases and nucleic acid translocases. Annu. Rev. Biochem. 76, 23-50 (2007).

66. Z. Yu, S. Y. Mersaoui, L. Guitton-Sert et al., DDX5 resolves R-loops at DNA double-strand breaks to promote DNA repair and avoid chromosomal deletions. NAR Cancer 2, zcaa028 (2020).

67. S. C. Sridhara, S. Carvalho, A. R. Grosso et al., Transcription dynamics prevent RNA-Mediated genomic instability through SRPK2-dependent DDX23 phosphorylation. Cell Rep. 18, 334-343 (2017).

68. I. X. Wang, C. Grunseich, J. Fox et al., Human proteins that interact with RNA/DNA hybrids. Genome Res. 28, 1405-1414 (2018).

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

70. J. Sollier, K. A. Cimprich, Breaking bad: R-loops and genome integrity. Trends Cell Biol. 25, 514-522 (2015).

71. M. N. Gonzalez, D. Blears, J. Q. Svejstrup, Causes and consequences of RNA polymerase II stalling during transcript elongation. Nat. Rev. Mol. Cell Biol. 22, 3-21 (2021).

72. C. Niehrs, B. Luke, Regulatory R-loops as facilitators of gene expression and genome stability. Nat. Rev. Mol. Cell Biol. 21, 167-178 (2020).

73. E. Hatchi, K. Skourti-Stathaki, S. Ventz et al., BRCA1 Recruitment to Transcriptional Pause Sites Is Required for R-Loop-Driven DNA Damage Repair. Mol. Cell 57, 636-647 (2015).

74. N. Kim, S. Jinks-Robertson, Transcription as a source of genome instability. Nat. Rev. Genet. 13, 204-214 (2012).

75. T. Aparicio, D. Megias, J. Mendez, Visualization of the MCM DNA helicase at replication factories before the onset of DNA synthesis. Chromosoma 121, 499-507 (2012).

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

77. T. Wu, R. Lyu, C. He, spKAS-seq reveals R-loop dynamics using low-input materials by detecting single-stranded DNA with strand specificity. Sci Adv 8, eabq2166 (2022).

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

79. P. Sarkies, C. Reams, L. J. Simpson, J. E. Sale, Epigenetic instability due to defective replication of structured DNA. Mol. Cell 40, 703-713 (2010).

80. C. M. Wickramasinghe, H. Arzouk, A. Frey, A. Maiter, J. E. Sale, Contributions of the specialised DNA polymerases to replication of structured DNA. DNA Repair 29, 83-90 (2015).

81. E. Walsh, X. X. Wang, M. Y. Lee, K. A. Eckert, Mechanism of replicative DNA polymerase d pausing and a potential role for DNA polymerase k in common fragile site replication. J. Mol. Biol. 425, 232-243 (2013).

82. P. Mukherjee, I. Lahiri, J. D. Pata, Human polymerase k uses a template-slippage deletion mechanism, but can realign the slipped strands to favour base substitution mutations over deletions. Nucleic Acids Res. 41, 5024-5035 (2013).

83. J. L. Sparks, H. Chon, S. M. Cerritelli et al., RNase H2-initiated ribonucleotide excision repair. Mol. Cell 47, 980-986 (2012).

84. Y. Wang, B. R. Knudsen, L. Bjergbaek, O. Westergaard, A. H. Andersen, Stimulated activity of human topoisomerases IIa and IIb on RNA-containing substrates. J. Biol. Chem. 274, 22839-22846 (1999).

85. J. S. Williams, D. J. Smith, L. Marjavaara et al., Topoisomerase 1-mediated removal of ribonucleotides from nascent leading-strand DNA. Mol. Cell 49, 1010-1015 (2013).

Chapter 4

N^2-alkyl-dG lesions trigger R-loop formation in genome

Introduction

Alkylating agents are produced from cell metabolism and are widely found in environmental pollutants (1-3). Constant exposure to those agents induces DNA damage in living cells, and the ensuing DNA adducts may result in genomic instability by impeding DNA replication and transcription (4, 5). A number of alkylating agents can react with the N^2 of 2′-deoxyguanosine (dG) in DNA to form N^2-alkyl-dG lesions in the genome. For example, N^2-(1-carboxyethyl)-2′-deoxyguanosine (N^2-CE-dG) is the major stable adduct formed in calf thymus DNA upon exposure to methylglyoxal (MG), a by-product of glycolysis, at physiological concentration and temperature (6-8). Benzo[*a*]pyrene-7,8-dihydrodiol-9,10-epoxide (BPDE), formed from metabolic activation of benzo[*a*]pyrene, reacts predominantly with the N^2 of dG (9, 10). In addition, the N^2 of dG is susceptible to reaction with formaldehyde, which could be induced endogenously, and acetaldehyde, which can be produced from ethanol metabolism or lipid peroxidation, and is also present in external sources, including diesel exhaust, cigarette smoke, etc. (11-13).

R-loops are non-B nucleic acid structures that emerge during transcription when the nascent RNA anneals with the template DNA strand, with a displaced non-template DNA strand (14). Initially thought to be rare byproducts of transcription, R-loops are now known to occur widely in bacteria, yeast, and higher eukaryotes throughout the cell cycle (15, 16). There are two types of R-loops, “scheduled” R-loops, which play roles in normal cellular processes, and “unscheduled” R-loops, formed during periods of cellular dysregulation and associated with

replication stress, DNA damage, and various human pathologies such as neurodegenerative disorders and cancer (17, 18). While R-loops are considered a potent source of DNA damage and genome instability, recent studies revealed that DNA double-strand breaks (DSBs) can induce R-loop formation (19-21). However, it remains unclear what the mechanisms are for their accumulation at DSBs and if other DNA lesions promote R-loop formation.

In light of the previous reports that N^2-alkyl-dG lesions can result in genomic instability by hindering DNA replication and transcription (22, 23), we hypothesized that these lesions, like DSBs, may also induce R-loop formation. In this study, we investigated the roles of N^2-alkyl-dG lesions in R-loop accumulation. Our results demonstrate, for the first time, that N^2-alkyl-dG lesions trigger R-loops formation, leading to genome instability in human cells.

Materials and Methods

Cell lines

HEK293T cells were purchased from ATCC. All cell lines used in this study were tested to be free of mycoplasma contamination using LookOut Mycoplasma PCR Kit (Sigma, MP0035). Cells were maintained in DMEM (Life Technologies) supplemented with 10% FBS (Invitrogen) and 1% penicillin/streptomycin (v/v) at 37°C in a humidified incubator with 5% CO_2. For genomic DNA extraction, cells were harvested at 80% confluency. DDX23 knockout cells in HEK293T background were generated previously using the CRISPR-Cas9 method. [D_9]-N^2-*n*Bu-dG and N^2-alkyl-dG-bearing DNA oligomers were synthesized following previously published procedures (4, 5, 24).

Synthesis and purification of N^2-heptynyl-2′-deoxyguanosine (N^2-heptynyl-dG)

N^2-heptynyl-dG was synthesized from 2-fluoro-6-*O*-(2-(4-nitrophenyl)ethyl)-2′-deoxyinosine (LGC Biosearch technologies, Inc.) (Figure 4-1). Briefly, 2-fluoro-6-*O*-(2-(4-nitrophenyl)ethyl)-2′-deoxyinosine (4.0 mg, 0.01 mmol) was dissolved in 80 μL dimethyl sulfoxide (DMSO), to which solution was added 6-heptyn-1-amine•HCl (7 mg, 0.05 mmol) and anhydrous *N*,*N*-diisopropylethylamine (DIEA, 3.8 μL, 0.02 mmol). The reaction mixture was stirred at 55°C for 3 days. The resulting mixture was dried *in vacuo*, and reconstituted in 1 M 1,8-diazabicyclo[5.4.0]undec-7-ene (DBU) (100 μL) to deprotect the nitrophenylethyl group. The crude product was dried *in vacuo* and redissolved in 1 mL doubly distilled water.

A Kinetex XB-C18 column (4.6 × 150 mm, 5 μm in particle size and 100 Å in pore size; Phenomenex Inc., Torrance, CA, USA) was employed to purify N^2-heptynyl-dG. Doubly distilled water and methanol were selected as mobile phases A and B, respectively. The gradient compromised of 10% B at 0-2 min and 10-80% B at 2-75 min, with the flow rate being 2.5 mL/min. The purified N^2-heptynyl-dG was confirmed by ESI-MS and MS/MS analyses (Figure 4-2).

Incorporation of N^2-*n*Bu-dG and N^2-heptynyl-dG into genomic DNA

HEK293T cells, and the isogenic *DDX23*$^{-/-}$, *POLK*$^{-/-}$ cells were seeded in 12-well plates with coverslips or 6-well plates at 37°C in a 5% CO_2 atmosphere. N^2-*n*Bu-dG or N^2-heptynyl-dG were added to the culture medium at a final concentration of 10 μM. After a 3-h incubation, the cells were harvested immediately, or after a 3- or 8-h incubation in fresh media without the modified nucleoside.

Enzymatic digestion and LC–MS/MS analysis of N^2-heptynyl-dG in cellular DNA

Genomic DNA was extracted from cells using Qiagen DNeasy Blood & Tissue Kit, and approximately 6 μg of DNA was recovered from a single well of cells. The extracted genomic DNA was subjected to enzymatic digestion following previously published procedures (25, 26). In brief, 1.0 μg of cellular DNA was digested with 10 units of nuclease P1 and 0.00125 unit of phosphodiesterase II in a buffer with 30 mM sodium acetate (pH 5.6), 1 mM $ZnCl_2$, and 2.5 nmol of *erythro*-9-(2-hydroxy-3-nonyl)adenine (EHNA, adenosine deaminase inhibitor). The above mixture was incubated at 37 °C for 24 h. After then, 1.0 unit of alkaline phosphatase, 0.0025 unit of phosphodiesterase I and one tenth volume of 0.5 M Tris-HCl (pH 8.9) were added. The resulting mixture was incubated at 37 °C for another 4 h and subsequently neutralized with 1.0 M formic acid. The enzymes in the digestion mixture were then removed by chloroform extraction. The aqueous phase was dried *in vacuo* and the resulting residues redissolved in water for LC-MS/MS analysis, where an LTQ linear ion-trap mass spectrometer (Thermo Electron, San Jose, CA, USA) and an Agilent Zorbax SB-C18 column (0.5 × 150 mm, 5 μm in particle size) were used. Mobile phases A and B were composed of 2 mM ammonia bicarbonate in doubly distilled water and acetonitrile, respectively. The gradient was 5 min of 5% mobile phase B, followed by 35 min of 5-90% mobile phase B at a flow rate of 8 μL/min. The temperature for the ion transfer tube was 300 °C, and the mass spectrometer was set up to acquire the full-scan MS/MS for the $[M+H]^+$ ions of N^2-heptynyl-dG and $[D_9]$-N^2-*n*Bu-dG.

Immunofluorescence microscopy

Immunofluorescence microscopy analysis of R-loop structure was conducted following previously published procedures (27). HEK293T cells and isogenic cells deficient of *DDX23* grown on coverslips were fixed with chilled methanol at room temperature for 15 min. The cells were then washed with PBS-TX (PBS containing 0.1% Triton X-100) and blocked with an antibody dilution solution (AbDil-Tx, PBS containing 2% BSA, 0.05% NaN_3) at 4°C for 6 h. The

cells were incubated with recombinant GFP-tagged catalytically dead RNase H1 (GFP-dRNH1) at 37°C for 1.5 h and then washed with PBS-TX. The samples were incubated with 4'-6-diamidino-2-phenylindole (DAPI) (Sigma) for 5 min to stain the DNA and the slides were mounted in Vectashield (Vector Laboratories, Burlingame, CA).

For imaging γ-H2AX, the cells were fixed and blocked following the same procedures as described above. The cells were subsequently incubated with the Phospho-Histone H2A.X (Ser139) antibody (2577S, Cell Signaling) in AbDil-Tx at room temperature for 1.5 h and then with Alexa Fluor 594 goat anti-rabbit IgG (Thermo Fisher). The samples were incubated with DAPI for 5 min to stain the DNA and the slides were mounted in Vectashield. Images were taken using a Zeiss 880 Inverted (Carl Zeiss, Oberkochen, Germany) confocal microscope with a 40x/1.4 oil immersion and quantified with ImageJ.

Western Blot

Western blot was conducted following previously published procedures with minor modifications, where protein samples were separated on an SDS-PAGE gel and transferred to a nitrocellulose membrane (Bio-Rad). After blocking with a blotting-grade blocker (Bio-Rad), the membrane was incubated in a solution containing primary antibody and 5% BSA (w/v) at 4°C overnight, and then incubated with the HRP-conjugated secondary antibody in a 5% blotting-grade blocker (w/v) at room temperature for 1 h. The immunoblots were detected using ECL Western blotting detection reagent (Amersham). Primary antibodies used in this study included α-Tubulin Antibody (DM1A) (Santa Cruz, sc-32293; 1:2000) and Phospho-Histone H2A.X (Ser139) Antibody (2577S, Cell Signaling Technology; 1:1000).

Construction of lesion-containing plasmids

The methods for the construction of the lesion-containing plasmids were described in detail elsewhere (28), where 12-mer N^2-alkyl-dG-containing ODNs were employed as the lesion-containing insert. Briefly, the damage-free control vector was digested with Nt.BstNBI to nick the double-stranded parental vector. The 25mer ODN arising from Nt.BstNBI cleavage was subsequently removed by annealing with excess complementary 25mer ODN and by centrifugation using a 100 kDa cutoff centrifugal filter. The 12mer N^2-alkyl-dG-containing ODN (5′-ATGGCGXGCTAT-3′, X= N^2-alkyl-dG) was 5′-phosphorylated and annealed into the gap together with a 13mer 5′-phosphorylated lesion-free ODN (5′-TCGGGAGTCGATG-3′). T4 DNA ligase was then added to seal the gap. The fully ligated, supercoiled plasmid was isolated from the ligation mixture by using agarose gel electrophoresis.

Cellular transcription, RNA isolation, and RT-PCR

Cellular transcription and RNA isolation, RT-PCR amplification, polyacrylamide gel electrophoressis (PAGE), and LC-MS/MS analysis were carried out as previously described (28). Briefly, the lesion-containing or lesion-free plasmids were individually premixed with the competitor plasmid at a molar ratio of 3:1 (lesion or control/competitor) for transfection. HEK293T, *DDX23*$^{-/-}$ cells, and the latter complemented with DDX23 (1×10^5) were seeded in 24-well plates and cultured overnight at 37 °C in a 5% CO_2 atmosphere, followed by transfection with 50 ng of the mixed plasmids and 450 ng of carrier plasmid (self-ligated pGEM-T, Promega) using TransIT-2020 (Mirus Bio). The cells were harvested at 24 h, the transcripts of the mixed plasmids were isolated using Total RNA Kit I (Omega), and residual DNA in the mixture was removed with a DNA-free kit (Ambion). The transcripts of interest were reverse-transcribed and PCR-amplified, as described elsewhere (28).

Restriction digestion, polyacrylamide gel electrophoresis (PAGE), and LC-MS/MS analysis

For each sample, 150 ng of the above-mentioned RT-PCR product was incubated with 5 U NcoI and 1 U shrimp alkaline phosphatase (rSAP) in 10 μL of 1× NEB buffer 3.1 at 37 °C for 1 h. The enzymes were heat-inactivated by incubation at 70 °C for 20 min, and to the mixture were added 5 U T4 polynucleotide kinase and 1.66 pmol [γ-^{32}P]ATP to radiolabel the newly liberated 5′-terminus in the template strand. The resultant mixture was heated at 65°C for 20 min and further digested with 2 U SfaNI in 20 μL of 1× NEB buffer 3.1 at 37°C for 1.5 h. The reaction was terminated with 20 μL of formamide gel-loading buffer (2×), and the DNA mixture was resolved by using 30% native PAGE (acrylamide/bis-acrylamide = 19:1) and quantified by phosphorimager analysis. The intensities of the radiolabeled DNA bands were used to calculate the relative bypass efficiency (RBE) with the following equation: RBE (%) = (lesion signal/competitor signal)/(control signal/competitor signal) × 100%, where the competitor signal was employed as the internal standard.

LC-MS and MS/MS were used to identify unambiguously the transcription products arising from N^2-alkyl-dG-containing templates, similar to those described elsewhere (28). RT-PCR products were treated with 50 U NcoI and 20 U rSAP in 250 μL of NEB buffer 3.1 at 37°C for 2 h, followed by heating at 80°C for 20 min. To the resulting solution was added 50 U SfaNI, and the reaction mixture was incubated at 37°C for 2 h, followed by extraction with phenol/chloroform/isoamyl alcohol (25:24:1, v/v). The aqueous phase was collected, to which were added 0.1 volume of 3.0 M sodium acetate and 2.5 volumes of ethanol to precipitate the DNA. The DNA pellet was reconstituted in water and subjected to LC–MS/MS analysis. An LTQ linear ion-trap mass spectrometer (Thermo Fisher Scientific) was set up for monitoring the fragmentations of the $[M - 3H]^{3-}$ ions of the 13-mer ODNs.

Chromatin immunoprecipitation-quantitative PCR (ChIP-qPCR) analysis of R-loop on lesion-bearing vector plasmids

R-ChIP-qPCR experiments were performed on cells expressing the D210N mutant of V5-tagged RNaseH1 protein following previously published procedures (29). Briefly, cells were seeded in 6-well plates at 37°C in a 5% CO_2 atmosphere. Lesion-containing or lesion-free plasmids (100 ng) were individually premixed with 900 ng carrier plasmid for transfection. After incubating for 6 h, the cells were washed with cold PBS and crosslinked with 1% formaldehyde at room temperature for 10-15 min. Fixation was stopped by adding glycine at a final concentration of 125 mM and incubation at room temperature for 15 min. After washing the plates twice with PBS, the cells were scraped off and the nuclei were extracted with cell lysis buffer (10 mM Tris-HCl, pH 8.0, 10 mM NaCl, 0.5% NP-40, and 1× protease inhibitor cocktail), and then suspended in nuclei lysis buffer (50 mM Tris-HCl pH 8.0, 10 mM EDTA, 1% SDS and 1 × protease inhibitor cocktail). Chromatin DNA was sheared to 250-600 bp in size by sonication. Approximately 5% chromatin fragment was saved as input and the remaining was incubated with magnetic beads conjugated with anti-V5 antibody at 4°C overnight. The beads were washed sequentially three times with washing buffer I (20 mM Tris-HCl, pH 8.0, 150 mM NaCl, 1% Triton X-100, 0.1% SDS, 2 mM EDTA and 1 × protease inhibitor cocktail), three times with washing buffer II (20 mM Tris-HCl, pH 8.0, 500 mM NaCl, 1% Triton X-100, 0.1% SDS, 2 mM EDTA and 1× protease inhibitor cocktail), once with washing buffer III (10 mM Tris-HCl, pH 8.0, 250 mM LiCl, 1% NP-40, 1% deoxycholate, 1 mM EDTA and 1 × protease inhibitor cocktail) and once with TE buffer (10 mM Tris-HCl pH 8.0 and 1 mM EDTA). The protein-chromatin complex was eluted with elution buffer (10 mM Tris-HCl, pH 8.0, 1% SDS, and 1 mM EDTA) and reverse crosslinked by incubation at 65°C overnight. After sequential RNase A and Proteinase K treatment, the DNA was purified using Cycle Pure Kit (Omega). Quantitative PCR was performed using standard protocols with Luna® Universal qPCR Master Mix (NEB, M3003X). Analysis was carried out using the $\Delta\Delta Ct$ method (fold enrichment) (30).

Strand-specific kethoxal-assisted single-stranded DNA (ssDNA) sequencing (spKAS-seq)

spKAS-seq was conduceted following previously published procedures (31) with minor modifications, where N^2-heptynyl-dG was added to the culture medium until its final concentration reached 10 μM. After a 3-h incubation, the cells were treated with 5 mM N_3-kethoxal dissolved in the culture medium for 10 min at 37°C, then harvested for genomic DNA isolation using the PureLink Genomic DNA Mini Kit (Thermo Fisher Scientific, K182002). The purified genomic DNA was subsequently mixed with 5 μl of 20 mM dibenzocyclooctyne-PEG_4-biotin conjugate (Sigma-Aldrich, 760749) and 10 μl of 10× phosphate-buffered saline (PBS), and the total volume was adjusted to 100 μl with 25 mM K_3BO_3. The resulting mixture was shaken gently at 37°C for 1.5 h, followed by addition of 5 μl of 10 mg/mL RNase A (Thermo Fisher Scientific, 12091039), and the mixture was shaken at 37°C for another 5 min. After the reaction, the DNA was isolated using the DNA Clean & Concentrator-5 kit (Zymo Research, D4013) and fragmented by sonication to 150 to 350 base pairs (bp).

A portion (~5%) of the above sonicated DNA was saved as input, and the rest was employed for enrichment using 10 μl of Dynabeads MyOne Streptavidin C1 (Thermo Fisher Scientific, 65001). The beads were washed with 1× binding and washing (B&W) buffer (5 mM tris-HCl, pH 7.4, 0.5 mM EDTA, 1 M NaCl, and 0.05% Tween 20), resuspended in 95 μl of 2× B&W buffer, and mixed with the sonicated DNA. The suspension was incubated at room temperature for 15 min. The beads were then washed once with 1× B&W buffer, twice with 100 mM NaOH solution to denature the dsDNA and remove the DNA strands that are not labeled by N_3-kethoxal, and once again with 1× B&W buffer. The DNA was eluted from the beads by heating the beads in 10 μl H_2O at 95°C for 10 min. The ensuing enriched DNA and the corresponding input were employed for library construction using the Accel-NGS Methyl-Seq DNA Library Kit (Swift, 30024).

spKAS-seq data processing

The sequencing reads of spKAS-seq data were checked with FastQC. Trim Galore was used to remove low-qualified bases and adapter-containing reads from raw spKAS-seq data. Trimmed reads were aligned to the human hg19 reference genome using Bowtie2 with the default configuration (32). Sam files were subsequently converted and sorted to binary alignment map (BAM) files using samtools sort (33). Duplicated reads were removed using Picard MarkDuplicates. Browser extended data (BED) files were converted to BedGraph files using bedtools genomecov (34). BedGraph files were then converted to BigWig files using bedGraphToBigWig from University of California, Santa Cruz (UCSC) precompiled utilities. All the metagene profile plots and heatmaps were generated using deepTools plotProfile and plotHeatmap (35). Definitions of R-loops by spKAS-seq were performed by following previously reported procedures (31). Two spKAS-seq replicates were used for R-loop identification. Genomic annotations were performed using Homer (36). The intersection between bed files was performed using BEDTools (37). The sequencing data generated in this study have been deposited into the NCBI Gene Expression Omnibus (GEO) under accession number GSExxx.

Clonogenic survival assay

Clonogenic survival assay was performed as described previously (38, 39). Briefly, parental and *DDX23*$^{-/-}$ HEK293T cells were plated in triplicate in six-well plates at a concentration of 150 cells per well, Cells were allowed to adhere to the plates for 8 h, and subsequently treated with BPDE (0, 0.25, 0.50, 0.75, 1.0, 1.5 μM) (Toronto research chemicals Inc, Canada) for 9 days. Cell colonies were fixed and stained in an aqueous solution containing 6.0% glutaraldehyde and 0.5% crystal violet. The plates were then rinsed with water and dried at room temperature in air.

Colonies were subsequently counted, and survival fraction (SF) was calculated using the following equation: SF = ($N_{\text{clonies, BPDE treated}}$) / ($N_{\text{clonies, control}}$).

Results

Minor-groove N^2-alkyl-dG lesions are known to impede transcription (4). We set out to examine if induction of these lesions could result in increased R-loops in cells. To this end, we first assessed if BPDE-induced DNA damage can trigger R-loop formation by using GFP-dRNH1, which can be used to detect R-loop structures (27). We observed a significant increase in R-loop level after exposure of cells to 2 μM BPDE for 3 h (Figure 4-3a, b). DDX23 is a helicase that functions in unwinding R-loops (40, 41). We observed that R-loops are substantially accumulated in DDX23-depleted cells. Notably, the increased immunoflourescence intensity in DDX23-deficient cells was attenuated upon treatment with recombinant RNase H, suggesting that these signals are derived from R-loops.

Next, we sought to determine if other N^2-alkyl-dG lesions could also induce R-loop formation in chromatin. Spratt and colleagues (42) showed that incubating cultured human cells with N^2-*p*-ethynylbenzyl-dG (N^2-EBn-dG) can allow for facile incorporation of the modified nucleoside into genomic DNA. Hence, we cultured HEK293T cells in a medium containing 10 μM N^2-*n*Bu-dG for 3 h, which facilitated the incorporation of approximately 15 N^2-*n*Bu-dG per 10^6 nucleosides into genomic DNA (4), and assessed the levels of R-loops in cells. We found a significant increase in R-loop formation at 3 h following the treatment (Figure 4-3c, d). A previous study revealed that nucleotide excision repair (NER) exerts moderate effects on the removal of N^2-*n*BudG from genomic DNA, and N^2-*n*Bu-dG incorporated in genomic DNA showed a time-dependent decrease after a 3-h repair (4). Hence, we monitored R-loop levels at different time points after removal of N^2-*n*Bu-dG from the culture medium. We found that the R-

loop level was significantly decreased after a 3-h repair compared to 0-h. An 8-h repair led to a slight decrease in R-loop compared to 3-h repair, which is in line with the previous finding that most N^2-*n*Bu-dG was repaired within 3 h (43). Notably, even after an 8-h repair, R-loops did not restore to the level observed for untreated cells, suggesting a delayed resolution of R-loops following N^2-*n*Bu-dG repair. Together, the R-loops are induced by N^2-*n*Bu-dG in genomic DNA and they can be resolved after lesion repair.

To validate that N^2-alkyl-dG induces R-loop formation in chromatin, we also performed spKAS-seq analysis (31) to explore the genome-wide distributions of R-loops in HEK293T cells with or without incorporation of N^2-heptynyl-dG into genomic DNA (Figures 4-11, and 4-12). Gene annotation analysis showed that 40.4% and 26.1% of R-loop peaks are located in promoter regions and gene bodies, respectively (Figure 4-4b), which is in keeping with the previously reported results (31). Meanwhile, after N^2-heptynyl-dG treatment, 41.5% and 26.4% of R-loop peaks are detected in the promoter region and gene bodies, which is quite similar to untreated cells. Surprisingly, R-loop peak numbers only increased slightly from 10231 to 11085 after N^2-heptynyl-dG treatment (Figure 4-4a). However, when we compared the intensity of those peaks commonly identified in N^2-heptynyl-dG-treated and untreated cells, the R-loop signal exhibits significant higher intensities in the treated cells (Figure 4-4c). Integrative Genomics Viewer (IGV) plots also showed a higher signal intensity in N^2-heptynyl-dG-treated cells compared to the untreated control. These results suggest that N^2-heptynyl-dG lesion induces the accumulation of R-loop structures in those regions that are susceptible to R-loop formation but does not substantially change R-loop distribution at the genome-wide scale. Our sequencing results are in line with the previous findings that DSB-induced R-loop formation is not a ubiquitous feature but rather is restricted to DSBs falling within regions that are transcriptionally active before DNA

damage (44, 45). These results substantiate our notion that N^2-alkyl-dG lesion triggers R-loop accumulation.

To explore whether the increased R-loop signal was indeed induced by N^2-*n*Bu-dG, we constructed plasmids carrying an N^2-alkyl-dG lesion downstream of transcription starting site by following our previously reported procedures (28). In this vein, we generated the plasmids containing N^2-*n*Bu-dG as well as *R* and *S* diastereomers of N^2-CE-dG (Figure 4-5a), which is a common N^2-alkyl-dG lesion induced by endogenous metabolism (46, 47). We then conducted R-ChIP-qPCR analysis for cells transfected with the N^2-alkyl-dG-bearing plasmids and lesion-free plasmid (28, 29). The recovery rate of the transcription starting site was significantly higher than that of the transcription termination site, suggesting R-loop formation at these loci (Figure 4-5b). All the N^2-alkyl-dG-containing plasmids exhibited around a 2-fold recovery rate of R-loop compared to the damage-free control plasmid, revealing that R-loop accumulation arises from the N^2-alkyl-dG lesions. Together, our results demonstrated that N^2-alkyl-dG lesions induce R-loop accumulation in an episomal plasmid.

We next asked whether the N^2-alky-dG-induced R-loops near the transcription starting site impede transcription. To this end, we conducted the competitive transcription and adduct bypass (CTAB) (28, 48) assay to assess quantitatively the transcription efficiencies in human cells. The CTAB method is based on RT-PCR amplification of RNA products emanating from concurrent transcription of lesion-bearing or lesion-free control plasmids together with a lesion-free competitor vector in cells (Figure 4-6). This is followed by polyacrylamide gel electrophoresis (PAGE) and LC-MS/MS analyses of short DNA fragments released from the restriction digestion of RT-PCR products (Figure 4-7a). Our results showed that the presence of N^2-alkyl-dG lesions on the transcribed strand led to pronouncedly diminished transcriptional bypass efficiencies in HEK293T cells, which is in line with a previous study (4, 28) showing that

N^2-alkyl-dG lesions impede transcription (Figure 4-7c). In addition, genetic ablation of DDX23 in HEK293T cells led to stronger diminutions in transcription efficiencies of N^2-alkyl-dG lesions compared to the parental cells, suggesting that R-loops impede transcription elongation. Moreover, complementation of *DDX23*$^{-/-}$ cells with wild-type DDX23 partially rescues the transcription suppression elicited by N^2-alkyl-dG lesions. In addition, PAGE and LC-MS and MS/MS analyses revealed that ablation of DDX23 does not lead to ribonucleotide misinsertions opposite the N^2-alkyl-dG lesions (Figure 4-7b, 4-8, and 4-9). These findings suggest that the N^2-alkyl-dG lesions trigger R-loop accumulation, thereby impeding transcription elongation.

Next, we sought to test whether DNA damage-induced R-loops contribute to genome instability. In replicating cells, R-loops are well known to impede the progression of replication forks and cause DNA DSB formation (49, 50). γH2AX is a sensitive molecular marker of DNA damage and accumulates at sites of DSBs (51, 52). We observed that γH2AX IF signal intensity is significantly higher in *DDX23*$^{-/-}$ cells compared to parental HEK293T cells, suggesting that R-loop accumulation may induce DSBs (Figure 4-10a, b). In addition, we observed increased γH2AX IF signal intensity in N^2-*n*Bu-dG-treated cells compared to the untreated control, where the R-loop signal was further increased in *DDX23*$^{-/-}$ cells. Our western blot results for γH2AX are in agreement with the IF assay results (Figure 4-10c). Together, our findings indicate that N^2-alkyl-dG lesions induce R-loop accumulation, which may further increase genome instability.

Finally, we conducted a clonogenic survival assay to examine how R-loop accumulation modulates the cellular sensitivity toward BPDE. As expected, BPDE treatment alone had a strong influence on impeding colony formation (Figure 4-10d). A concentration of 0.75 μM BPDE reduced the number of colonies by 80–90%, while a concentration of 1.5 μM completely inhibited colony growth. Genetic ablation of DDX23 strongly sensitized cells to BPDE, and

fewer colonies were formed from cells treated with different concentrations of BPDE compared to parental HEK293T cells.

Discussion

R-loops are dynamic nucleic acid structures occuring during transcription when the nascent RNA forms a hybrid duplex with the template DNA strand, leaving a displaced single-stranded DNA (53). R-loops are recognized as sources of DNA damage, leading to genome instability (14, 54), though the underlying mechanisms are still under investigation (55). Accordingly, unscheduled R-loops have various pathological consequences, including mutagenesis, replication stress, increased genomic instability, and promotion of cancer development (18). On the other hand, R-loops also play beneficial roles by regulating various processes, e.g., transcription, chromatin structure, DNA repair, and telomere maintenance in physiological contexts (14). Although R-loops are known to result in DNA damage, recent studies have surprisingly shown that DSBs induce the formation of R-loops (19-21). However, the mechanism leading to their accumulation at DSBs remains unclear. Additionally, it is unknown whether other DNA lesions can induce R-loop formation.

In this study, we demonstrated that N^2-alkyl-dG lesions elicit R-loop accumulation in human cells for the first time. Recently, NER pathway was shown to be responsible for removing adducts formed at the N^2 position of dG (43, 56). By using immunoflourescence assay, we showed that different N^2-alkyl-dG lesions, such as N^2-BPDE-dG and N^2-*n*Bu-dG, can induce the accumulation of R-loops in HEK293T cells. Upon repair of N^2-nBu-dG, the R-loop signals decreased but did not restore to the level observed in untreated levels, confirming that these R-loops are induced by the N^2-alkyl-dG and that R-loop resolution lags behind lesion repair.

To explore this effect on a genome-wide scale, we performed R-loop sequencing of HEK293T cells with or without N^2-heptynyl-dG treatment. Our results indicated an increase in R-loop signal intensity at those genomic regions already enriched with R-loops, whereas no significant induction of new R-loops in other regions were observed. These results suggest that N^2-alkyl-dG-induced R-loops formation is not a ubiquitous feature but rather is limited to N^2-alkyl-dG lesions formed in those regions that are transcriptionally active. Our results are in line with the previous findings that the accumulation of hybrids in cis to DSBs induced by enzymatic cleavage at loci occupied by RNAPII before damage but not around DSBs induced in intergenic loci despite equivalent levels of DSB induction (40, 45). These results support a "transcription regulation" model, where lesion-induced R-loops may form as a consequence of transcriptional repression occurring in cis to N^2-alkyl-dG lesion, rather than as a consequence of *de novo* transcription at lesion sites.

We further validated the N^2-alkyl-dG-induced R-loop accumulation by incorporating several different N^2-alkyl-dG lesions into the transcription start site of a plasmid. R-ChIP-qPCR results confirmed the R-loop formation at the transcription start site of the lesion-free plasmid. Moreover, R-loop signal is increased significantly in those N^2-alkyl-dG lesion-containing plasmids compared to the control, supporting that R-loop accumulation is directly induced by N^2-alkyl-dG lesions in human cells. These plasmids were also employed to assess the effects of N^2-alkyl-dG-induced R-loops in transcription efficiency by virtue of the CTAB assay. Our results showed that all N^2-alkyl-dG lesions tested impede transcription, which is consistent with previous studies (4, 24). Genetic depletion of DDX23, a helicase that unwinds R-loop structure, further attenuated transcription efficiency, indicating that R-loops impede transcription elongation. A hypothesis could be proposed: N^2-alkyl-dG lesions inhibit transcription and elongation, and induce R-loop formation, which further attenuates transcription and triggers genome instability.

This notion finds its support by a recent observation that R-loops contribute to polycomb repressive complex 1 recruitment and transcriptional repression of polycomb-repressed gene at the genome-wide scale (57).

We also explored the contribution of N^2-alkyl-dG lesion-induced R-loops to genome instability. Our immunoflourescence and Western blots results revealed an increased signal indicating the formation of DSBs after R-loop accumulation induced by N^2-alkyl-dG. These results are supported by a previous study indicating that unscheduled R-loops are actively processed into DSBs by NER endonucleases XPF and XPG (58). In addition, our clonogenic assay revealed that genetic ablation of DDX23, a R-loops helicase, exacerbated the cytotoxicity of the N^2-alkyl-dG.

In summary, we demonstrated that lesions formed at the N^2 position of dG induce R-loop accumulation on human chromatin. Our results suggest that N^2-alkyl-dG-induced R-loops further impede transcription and contribute to genome instability, which imply a potential therapeutical strategy through the combination of R-loops helicase inhibitors and DNA alkylating anticancer drugs. It would be intriguing to explore whether N^2-alkyl-dG lesion-induced R-loops facilitate the repair of these lesions by recruiting other repair proteins, and if other lesions can promote the formation of R-loops.

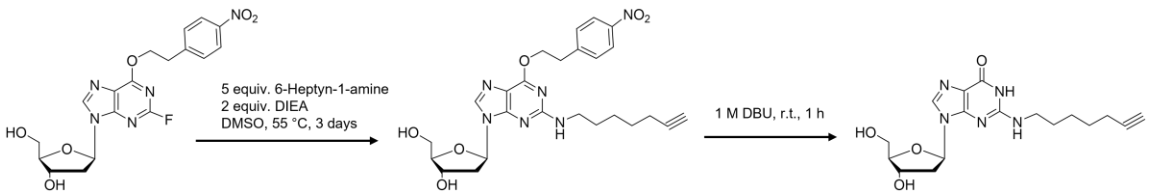

Figure 4-1. A Scheme showing the chemical synthesis of N^2-heptynyl-dG.

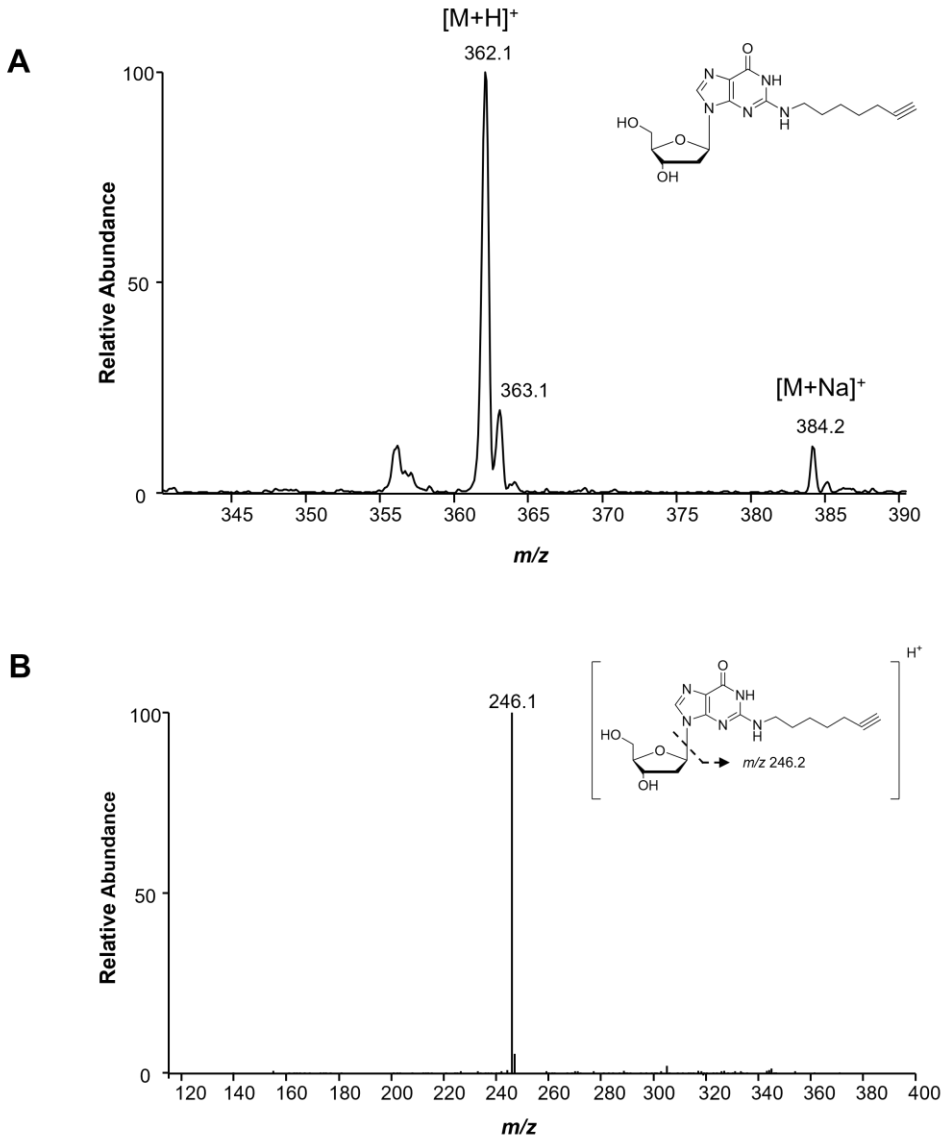

Figure 4-2. ESI-MS (A) and MS/MS (B) analysis of N^2-heptynyl-dG.

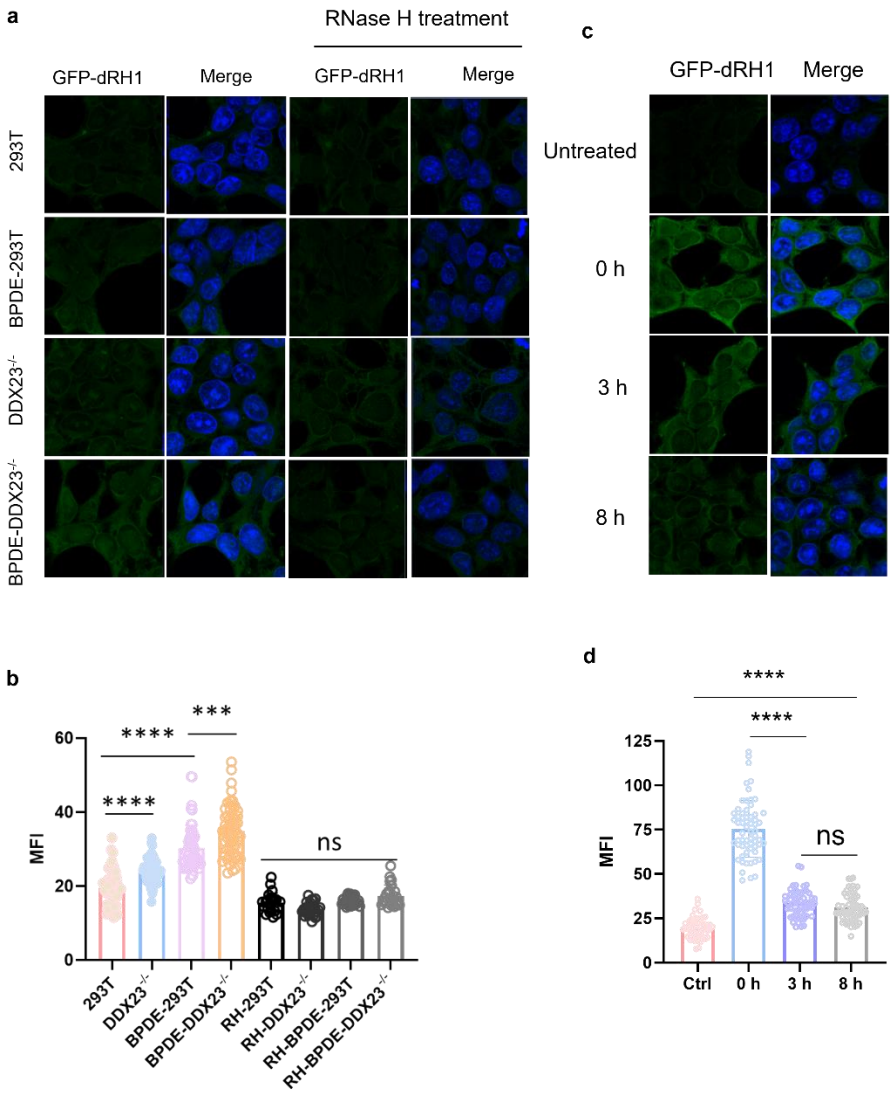

Figure 4-3. N^2-alkyl-dG lesions induce elevated R-loop formation. (a) Immunofluorescence microscopy analysis of R-loops in HEK293T and the isogenic *DDX23*$^{-/-}$ cells treated with or without BPDE. After fixation, coverslips were treated with or without RNase H. (b) Quantification of nuclear mean fluorescence intensities (MFI) for the conditions shown in a. (c) Immunofluorescence microscopy analysis of R-loops in cells with incorporation of N^2-*n*Bu-dG. After a 3-h exposure to 10 μM of N^2-*n*Bu-dG, the cells were harvested immediately, or cultured in fresh medium without the modified nucleoside for another 3 or 8 h. (d) Quantification of nuclear MFI for the conditions shown in c. The data represent the mean ± SD of results obtained from three biological replicates. ns, $p > 0.05$; ***, $p < 0.001$; ****, $p < 0.0001$.

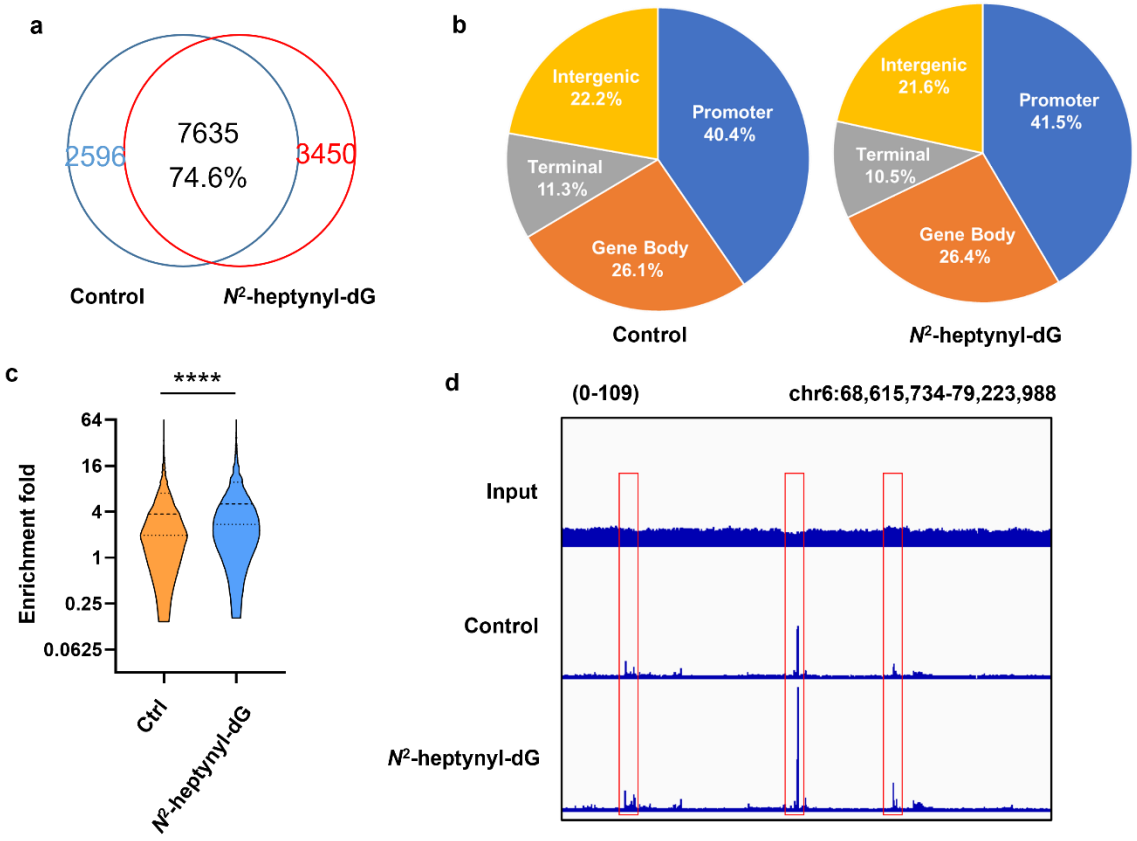

Figure 4-4. spKAS-seq revealed increased accumulation of R-loops in the N^2-heptynyl-dG treated HEK293T cells. (a) A Venn diagram depicting the overlap of R-loop peaks between control HEK293T cells and N^2-heptynyl-dG-treated HEK293T cells. (b) The genomic distribution of R-loop peaks mapped by sp-KAS-seq. Various genomic features are color-coded according to the labels at the bottom. (c) Fold enrichment of overlapped peaks between the N^2-heptynyl-dG-treated and untreated cell lines. (d) Genome browser view displaying the peak enrichment in N^2-heptynyl-dG-treated and control HEK293T cells.

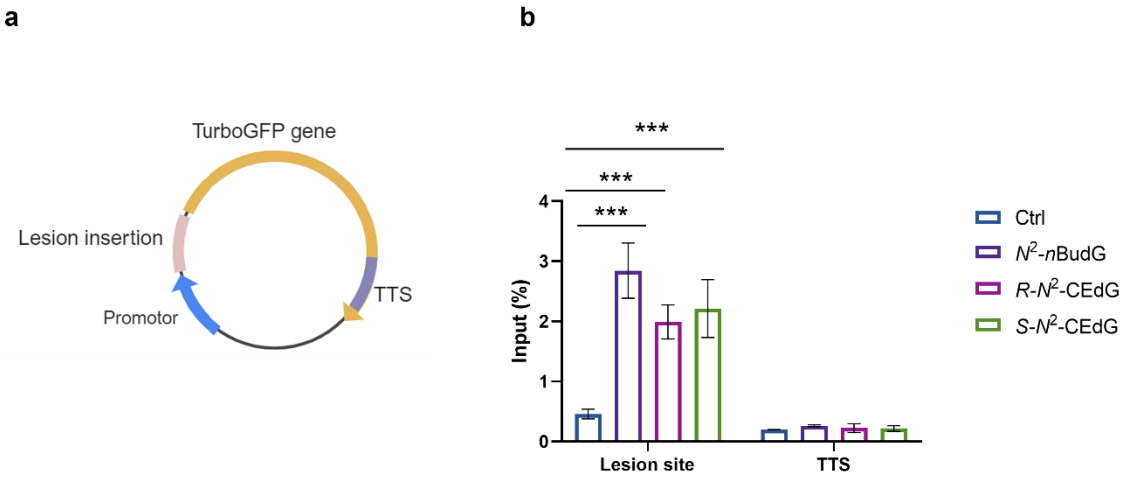

Figure 4-5. R-loops are accumulated at N^2-alkyl-dG incorporation site. (a) A schematic diagram showing the lesion-bearing vector, 'TTS' indicates the transcription terminal site (b) R-ChIP-qPCR analyses in the lesion insertion sites and TTSs. The data represent the mean ± S.D. of results obtained from two biological replicates, where each biological replicate was analyzed twice. ns, $p > 0.05$; ***, $p < 0.001$.

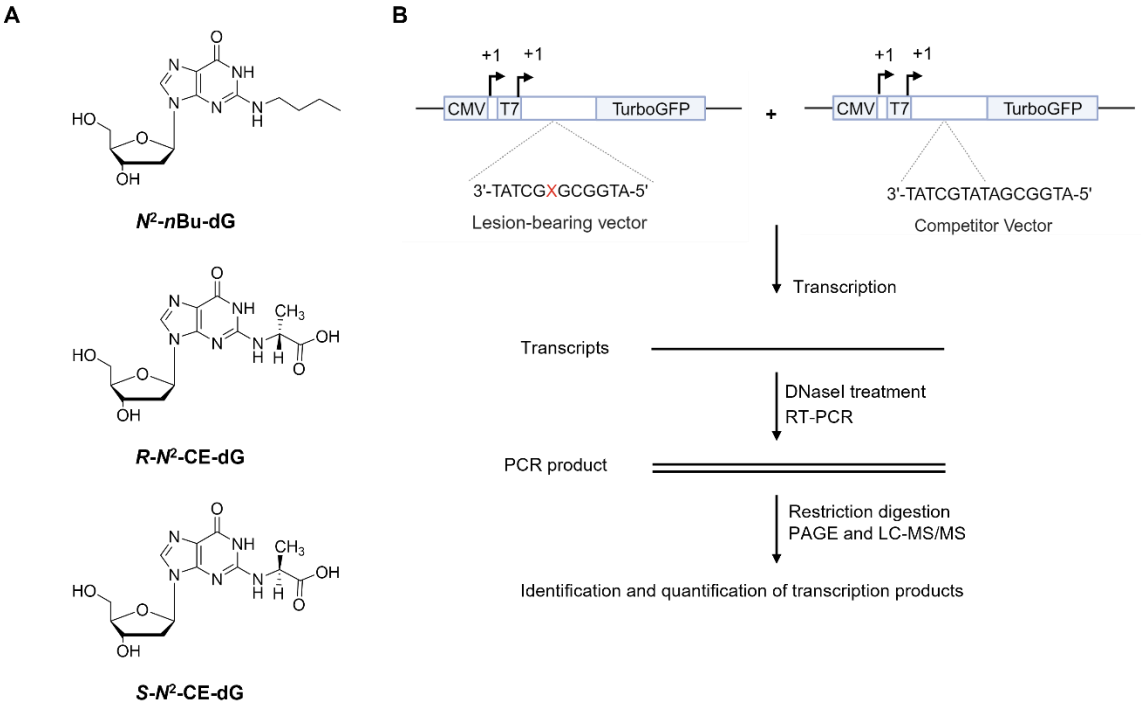

Figure 4-6. Scheme of CTAB assay. (A) N^2-alkyl-dG lesions investigated in CTAB assay. (B) A schematic diagram illustrating the CTAB assay system. 'X' indicates a lesion base or the corresponding unmodified base, which is located on the transcribed strand of the TurboGFP gene downstream of the CMV and T7 promoters. The arrowheads indicate the +1 transcription start sites of the CMV and T7 promoters.

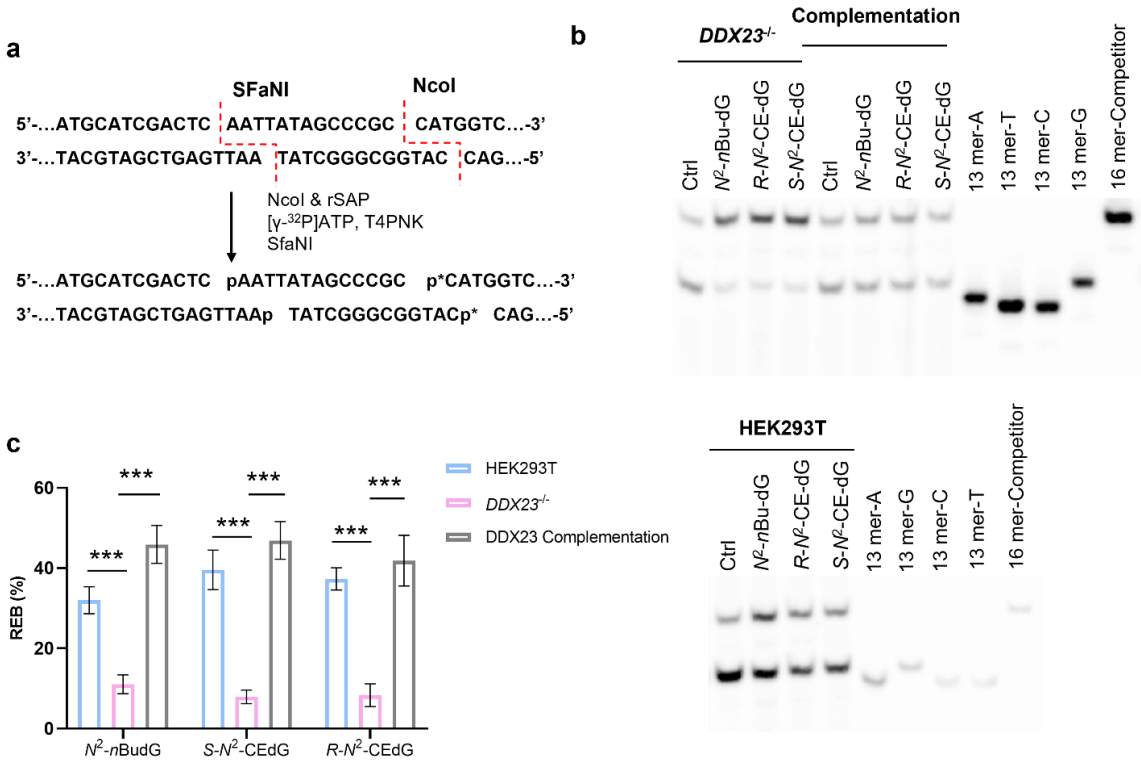

Figure 4-7. Restriction digestion and postlabeling method for determining the transcriptional bypass efficiencies of N^2-alkyl-dG lesions in HEK293T, *DDX23*$^{-/-}$, and the latter complemented with DDX23. (a) A schematic diagram depicting the selective labeling of the template strand via sequential digestion of the RT-PCR products. "p*" denotes a ^{32}P-labeled phosphate group. (b) Lesion-containing plasmids were individually premixed with the competitor plasmid at a molar ratio of 3:1 (lesion/competitor) for transfection into *DDX23*$^{-/-}$, DDX23 complementation cells and the transcripts were isolated from cells at 24 h following transfection. The synthetic ODNs are designated as "16mer", which has the same sequence as the restriction fragment from the competitor genome; "13 mer-G", "13 mer-A", "13 mer-C", "13 mer-T", which represent the restriction fragments with A, T, C, and G formed at the lesion site in the PCR product. (c) Relative transcriptional bypass efficiencies (RBEs) of N^2-alkyl-dG lesions. The data represent the mean ± S.D. of results from three independent experiments. ***, $p < 0.001$.

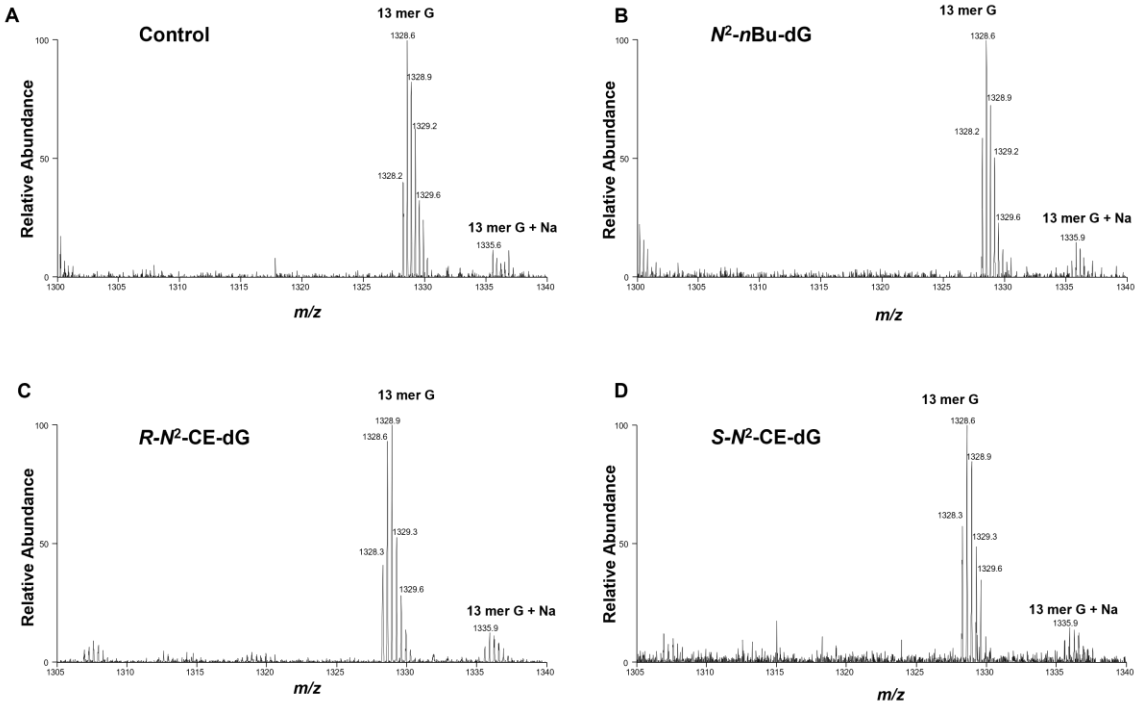

Figure 4-8. Higher resolution 'ultra zoom-scan' ESI-MS for monitoring the $[M-3H]^{3-}$ ions of the restriction fragments of interest from transcription of non-lesion control (A) and N^2-alkyl-dG lesions (B-D) in *DDX23*$^{-/-}$ cells.

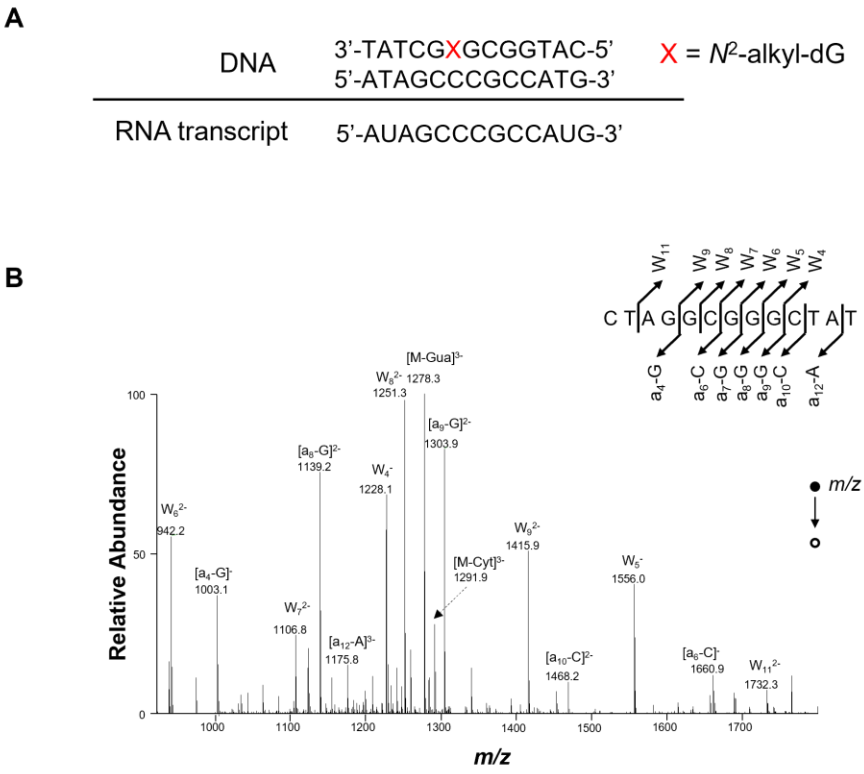

Figure 4-9. LC-MS/MS for monitoring the restriction digestion products of interest corresponding to wild-type or mutated transcripts arising from N^2-alkyl-dG containing substrates in HEK293T cells. (A)The sequence information of transcript is indicated below the double-stranded DNA construct. (B) The product-ion spectra of the ESI-produced $[M-3H]^{3-}$ ions (*m/z* 1328.6) of the 13-mer fragments of the wild-type sequence d(CATGGCGGGCTAT).

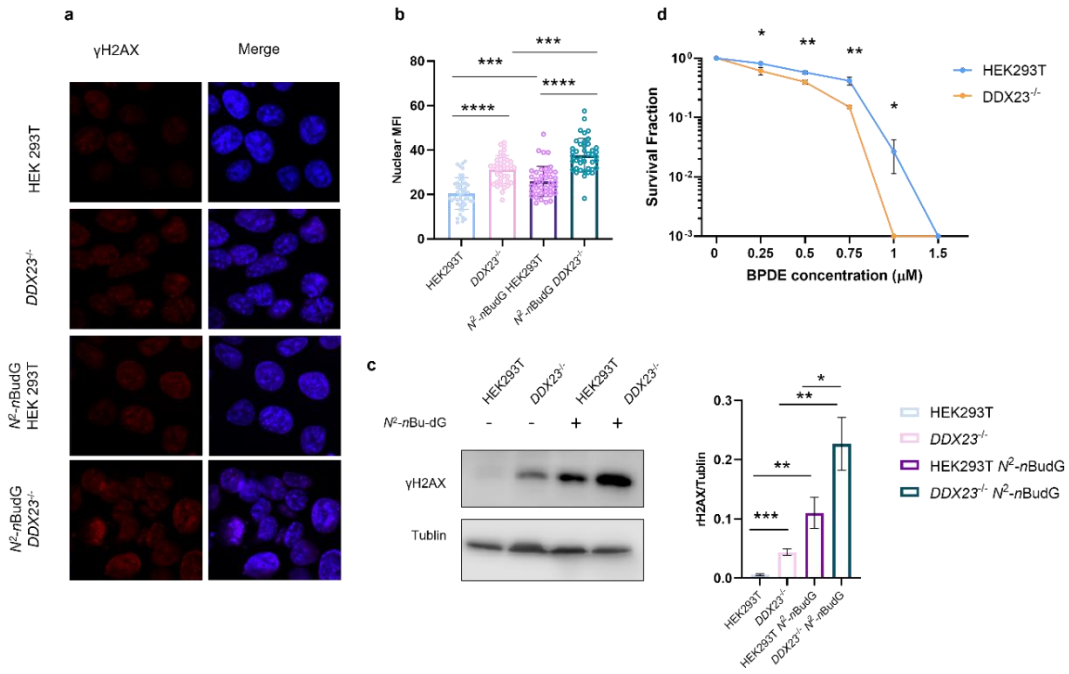

Figure 4-10. N^2-alkyl-dG incorporation induces genome instability. (a) Representative images of γ-H2AX foci in HEK293T and *DDX23*$^{-/-}$ cells treated with or without N^2-*n*Bu-dG. (b) Quantification of nuclear γ-H2AX mean fluorescence intensities (MFI) for the conditions shown in a. (c) γ-H2AX protein levels in whole-cell lysates of parental HEK293T and the isogenic *DDX23*$^{-/-}$ cells (n = 3). (d) Clonogenic cell survival assay results for HEK293T and the isogenic *DDX23*$^{-/-}$ cells upon BPDE treatment. Cells were plated, and the number of colonies was counted 9 days later. The surviving fractions (SF) were calculated after normalizing to the plating efficiency of the cells without BPDE treatment. The data represent the mean ± S.D. of results from three independent experiments. *, $0.01 < p < 0.05$; **, $0.001 < p < 0.01$.

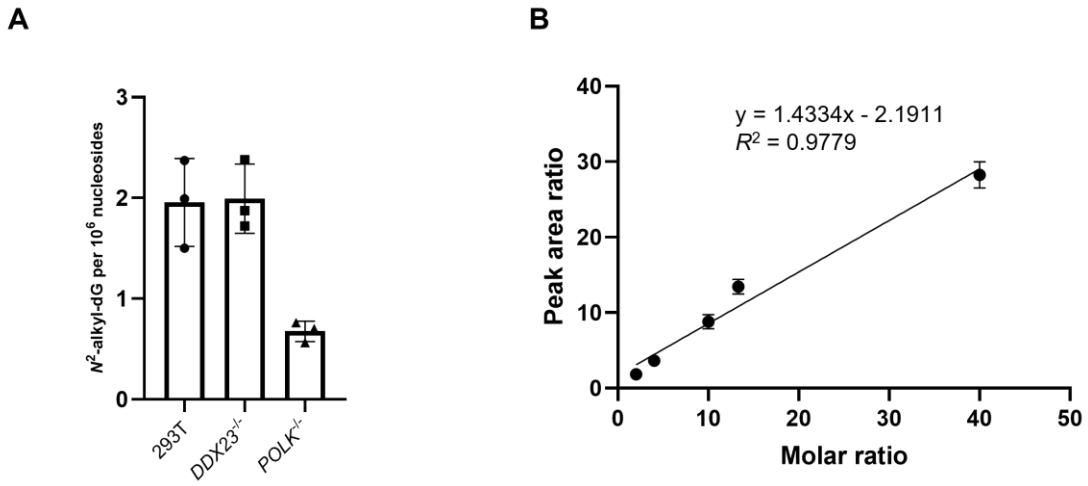

Figure 4-11. N^2-heptynyl-dG incorporation assay. (A) Frequencies of N^2-heptynyl-dG in cellular DNA isolated from parental and *DDX23*$^{-/-}$, *POLK*$^{-/-}$ HEK293T cells. All cells were exposed to 10 μM of N^2-heptynyl-dG for 3 h. (B) A calibration curve for the quantification of N^2-heptynyl-dG. The calibration curve was constructed by plotting the peak area ratios found in the selected-ion chromatograms for the N^2-heptynyl-dG over [D_9]-N^2-*n*Bu-dG vs. the molar ratios of the N^2-heptynyl-dG over [D_9]-N^2-*n*Bu-dG.

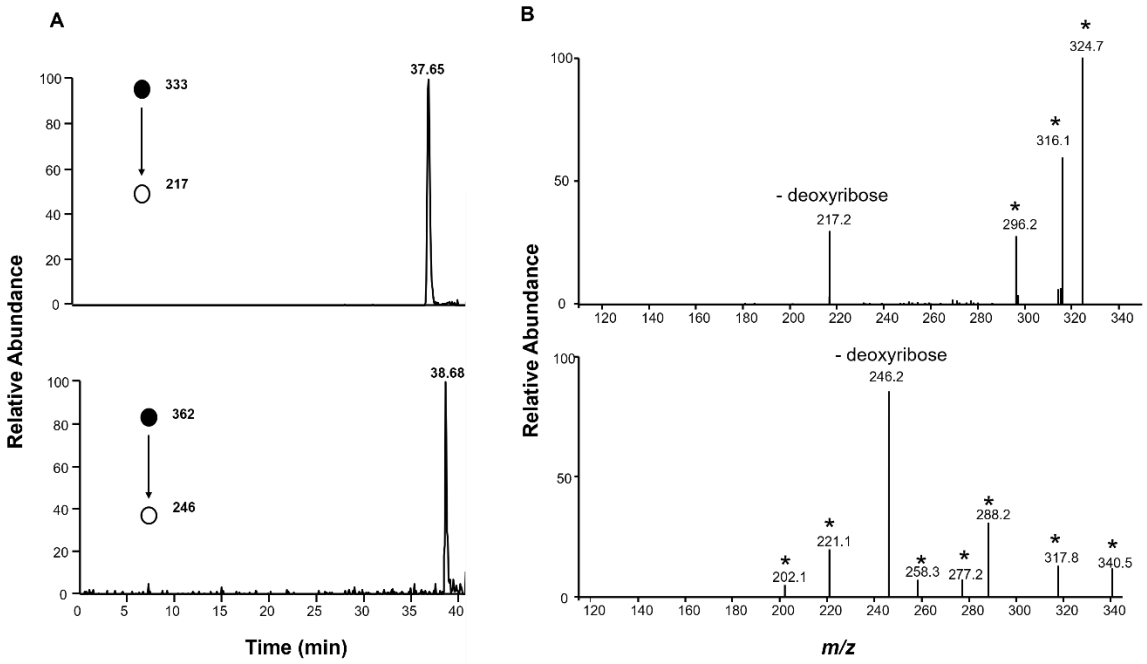

Figure 4-12. Representative selected-ion chromatograms (SICs) (A) and MS/MS (B) for monitoring m/z 333 to 217 (top panel) and 362 to 246 (bottom panel) transitions for the $[M + H]^+$ ions of $[D_9]$-N^2-*n*Bu-dG and N^2-heptynyl-dG. Ions labeled with "*" are due to the fragmentation of other co-eluting species.

Primer names	Sequences
Lesion insertion site qPCR forward	5′-CAGAGCTGGTTTAGTGAACCGT-3′
Lesion insertion site qPCR reverse	5′-GGTACCGTCGACTGCAGAAT-3′
Plasmid TTS qPCR forward	5′-AGGATCACAGCAACACCGAG-3′
Plasmid TTS qPCR reverse	5′-TTCTTCACCGGCATCTGCAT-3′
CTAB assay reverse transcription PCR	5′-TCGGTGTTGCTGTGAT-3′
CTAB assay amplification PCR forward	5′-CTAGCGGATGCATCGACTC-3′
CTAB assay amplification PCR reverse	5′-TGCTGCGGATGATCTTGTCG-3′

Table 4-1. Oligonucleotides used in this study.

References

1. D. Fu, J. A. Calvo, L. D. Samson, Balancing repair and tolerance of DNA damage caused by alkylating agents. Nat Rev Cancer 12, 104-120 (2012).

2. T. Helleday, DNA repair pathways as target for cancer therapy. Toxicology Letters 211, S9-S9 (2012).

3. B. Sedgwick, P. A. Bates, J. Paik, S. C. Jacobs, T. Lindahl, Repair of alkylated DNA: recent advances. DNA Repair (Amst) 6, 429-442 (2007).

4. Y. Tan, S. Guo, J. Wu et al., DNA Polymerase eta Promotes the Transcriptional Bypass of N(2)-Alkyl-2'-deoxyguanosine Adducts in Human Cells. J Am Chem Soc 143, 16197-16205 (2021).

5. J. Wu, H. Du, L. Li et al., The Impact of Minor-Groove N(2)-Alkyl-2'-deoxyguanosine Lesions on DNA Replication in Human Cells. ACS Chem Biol 14, 1708-1716 (2019).

6. P. J. Thornalley, The glyoxalase system: new developments towards functional characterization of a metabolic pathway fundamental to biological life. Biochem. J. 269, 1-11 (1990).

7. L. A. Fothergill-Gilmore, P. A. Michels, Evolution of glycolysis. Prog Biophys Mol Biol 59, 105-235 (1993).

8. P. J. Thornalley, Pharmacology of methylglyoxal: formation, modification of proteins and nucleic acids, and enzymatic detoxification--a role in pathogenesis and antiproliferative chemotherapy. Gen Pharmacol 27, 565-573 (1996).

9. M. Lodovici, C. Luceri, F. Guglielmi et al., Benzo(a)pyrene diolepoxide (BPDE)-DNA adduct levels in leukocytes of smokers in relation to polymorphism of CYP1A1, GSTM1, GSTP1, GSTT1, and mEH. Cancer Epidemiol Biomarkers Prev 13, 1342-1348 (2004).

10. H. Bartsch, DNA adducts in human carcinogenesis: etiological relevance and structure-activity relationship. Mutat Res 340, 67-79 (1996).

11. G. Cheng, Y. Shi, S. J. Sturla et al., Reactions of formaldehyde plus acetaldehyde with deoxyguanosine and DNA: formation of cyclic deoxyguanosine adducts and formaldehyde cross-links. Chem. Res. Toxicol. 16, 145-152 (2003).

12. L. Mulderrig, J. I. Garaycoechea, Z. K. Tuong et al., Aldehyde-driven transcriptional stress triggers an anorexic DNA damage response. Nature 600, 158-163 (2021).

13. L. B. Pontel, I. V. Rosado, G. Burgos-Barragan et al., Endogenous formaldehyde is a hematopoietic stem cell genotoxin and metabolic carcinogen. Mol. Cell 60, 177-188 (2015).

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

15. V. Bhatia, E. Herrera-Moyano, A. Aguilera, B. Gomez-Gonzalez, The Role of Replication-Associated Repair Factors on R-Loops. Genes (Basel) 8 (2017).

16. K. S. Lang, A. N. Hall, C. N. Merrikh et al., Replication-Transcription Conflicts Generate R-Loops that Orchestrate Bacterial Stress Survival and Pathogenesis. Cell 170, 787-799 e718 (2017).

17. P. A. Ginno, P. L. Lott, H. C. Christensen, I. Korf, F. Chedin, R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. Mol Cell 45, 814-825 (2012).

18. T. Garcia-Muse, A. Aguilera, R Loops: From Physiological to Pathological Roles. Cell 179, 604-618 (2019).

19. W. T. Lu, B. R. Hawley, G. L. Skalka et al., Drosha drives the formation of DNA:RNA hybrids around DNA break sites to facilitate DNA repair. Nat Commun 9, 532 (2018).

20. G. D'Alessandro, D. R. Whelan, S. M. Howard et al., BRCA2 controls DNA:RNA hybrid level at DSBs by mediating RNase H2 recruitment. Nat Commun 9, 5376 (2018).

21. C. Ohle, R. Tesorero, G. Schermann et al., Transient RNA-DNA Hybrids Are Required for Efficient Double-Strand Break Repair. Cell 167, 1001-1013 e1007 (2016).

22. A. Tubbs, A. Nussenzweig, Endogenous DNA Damage as a Source of Genomic Instability in Cancer. Cell 168, 644-656 (2017).

23. B. Yuan, J. Wang, H. Cao, R. Sun, Y. Wang, High-throughput analysis of the mutagenic and cytotoxic properties of DNA lesions by next-generation sequencing. Nucleic Acids Res 39, 5945-5954 (2011).

24. C. You, X. Dai, B. Yuan et al., A quantitative assay for assessing the effects of DNA lesions on transcription. Nat Chem Biol 8, 817-822 (2012).

25. S. Guo, J. Leng, Y. Tan, N. E. Price, Y. Wang, Quantification of DNA Lesions Induced by 4-(Methylnitrosamino)-1-(3-pyridyl)-1-butanol in Mammalian Cells. Chem Res Toxicol 32, 708-717 (2019).

26. J. Wang, B. Yuan, C. Guerrero et al., Quantification of oxidative DNA lesions in tissues of Long-Evans Cinnamon rats by capillary high-performance liquid chromatography-tandem mass spectrometry coupled with stable isotope-dilution method. Anal Chem 83, 2201-2209 (2011).

27. M. P. Crossley, J. R. Brickner, C. Song et al., Catalytically inactive, purified RNase H1: A specific and sensitive probe for RNA-DNA hybrid imaging. J. Cell Biol. 220 (2021).

28. C. You, Y. Wang, Quantitative measurement of transcriptional inhibition and mutagenesis induced by site-specifically incorporated DNA lesions in vitro and in vivo. Nat Protoc 10, 1389-1406 (2015).

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

30. F. Tang, Y. Wang, Z. Gao, S. Guo, Y. Wang, Polymerase eta Recruits DHX9 Helicase to Promote Replication across Guanine Quadruplex Structures. J Am Chem Soc 144, 14016-14020 (2022).

31. T. Wu, R. Lyu, C. He, spKAS-seq reveals R-loop dynamics using low-input materials by detecting single-stranded DNA with strand specificity. Sci Adv 8, eabq2166 (2022).

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

33. H. Li, B. Handsaker, A. Wysoker et al., The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078-2079 (2009).

34. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-842 (2010).

35. F. Ramírez, D. P. Ryan, B. Grüning et al., deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Research 44, W160-W165 (2016).

36. S. Heinz, C. Benner, N. Spann et al., Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell 38, 576-589 (2010).

37. A. R. Quinlan, I. M. Hall, BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26, 841-842 (2010).

38. N. A. Franken, H. M. Rodermond, J. Stap, J. Haveman, C. van Bree, Clonogenic assay of cells in vitro. Nat Protoc 1, 2315-2319 (2006).

39. W. Miao, D. Bade, Y. Wang, Targeted Proteomic Analysis Revealed Kinome Reprogramming during Acquisition of Radioresistance in Breast Cancer Cells. J Proteome Res 20, 2830-2838 (2021).

40. S. C. Sridhara, S. Carvalho, A. R. Grosso et al., Transcription Dynamics Prevent RNA-Mediated Genomic Instability through SRPK2-Dependent DDX23 Phosphorylation. Cell Reports 18, 334-343 (2017).

41. J. P. Matson, L. Zou, A genome-wide and cotranscriptional suppressor of R loops. Genes Dev 34, 863-864 (2020).

42. A. S. Gowda, M. Lee, T. E. Spratt, N(2) -Substituted 2'-Deoxyguanosine Triphosphate Derivatives as Selective Substrates for Human DNA Polymerase kappa. Angew Chem Int Ed Engl 56, 2628-2631 (2017).

43. S. Guo, L. Li, K. Yu, Y. Tan, Y. Wang, LC-MS/MS for Assessing the Incorporation and Repair of N(2)-Alkyl-2'-deoxyguanosine in Genomic DNA. Chem Res Toxicol 35, 1814-1820 (2022).

44. S. Cohen, N. Puget, Y. L. Lin et al., Senataxin resolves RNA:DNA hybrids forming at DNA double-strand breaks to prevent translocations. Nat Commun 9, 533 (2018).

45. A. S. Bader, M. Bushell, DNA:RNA hybrids form at DNA double-strand breaks in transcriptionally active loci. Cell Death Dis 11, 280 (2020).

46. B. Yuan, H. Cao, Y. Jiang, H. Hong, Y. Wang, Efficient and accurate bypass of N2-(1-carboxyethyl)-2'-deoxyguanosine by DinB DNA polymerase in vitro and in vivo. Proc Natl Acad Sci U S A 105, 8679-8684 (2008).

47. H. Cao, Y. Jiang, Y. Wang, Stereospecific synthesis and characterization of oligodeoxyribonucleotides containing an N2-(1-carboxyethyl)-2'-deoxyguanosine. J Am Chem Soc 129, 12123-12130 (2007).

48. C. You, P. Wang, X. Dai, Y. Wang, Transcriptional bypass of regioisomeric ethylated thymidine lesions by T7 RNA polymerase and human RNA polymerase II. Nucleic Acids Res 42, 13706-13713 (2014).

49. R. D. Paulsen, D. V. Soni, R. Wollman et al., A genome-wide siRNA screen reveals diverse cellular processes and pathways that mediate genome stability. Mol Cell 35, 228-239 (2009).

50. A. Helmrich, M. Ballarino, L. Tora, Collisions between replication and transcription complexes cause common fragile site instability at the longest human genes. Mol Cell 44, 966-977 (2011).

51. P. L. Collins, C. Purman, S. I. Porter et al., DNA double-strand breaks induce H2Ax phosphorylation domains in a contact-dependent manner. Nat Commun 11, 3158 (2020).

52. A. Kinner, W. Wu, C. Staudt, G. Iliakis, Gamma-H2AX in recognition and signaling of DNA double-strand breaks in the context of chromatin. Nucleic Acids Res 36, 5678-5694 (2008).

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

54. D. Zong, P. Oberdoerffer, P. J. Batista, A. Nussenzweig, RNA: a double-edged sword in genome maintenance. Nat Rev Genet 21, 651-670 (2020).

55. A. Promonet, I. Padioleau, Y. Liu et al., Topoisomerase 1 prevents replication stress at R-loop-enriched transcription termination sites. Nat Commun 11, 3940 (2020).

56. V. Shafirovich, M. Kolbanovskiy, K. Kropachev et al., Nucleotide Excision Repair and Impact of Site-Specific 5',8-Cyclopurine and Bulky DNA Lesions on the Physical Properties of Nucleosomes. Biochemistry 58, 561-574 (2019).

57. K. Skourti-Stathaki, E. Torlai Triglia, M. Warburton et al., R-Loops Enhance Polycomb Repression at a Subset of Developmental Regulator Genes. Mol Cell 73, 930-945 e934 (2019).

58. J. Sollier, C. T. Stork, M. L. Garcia-Rubio et al., Transcription-coupled nucleotide excision repair factors promote R-loop-induced genome instability. Mol Cell 56, 777-785 (2014).

Chapter 5

Concluding Remarks and Future Directions

Genome instability is a hallmark of most cancers. In hereditary cancers, genomic instability results from mutations in DNA repair genes and drives cancer development, as predicted by the mutator hypothesis. A high frequency of externally induced DNA damage can be one source of genome instability since DNA damage can lead to inaccurate translesion DNA synthesis past the damage or errors in repair, leading to mutations. Another source of genome instability refers to the high rate by which chromosome structure and number change over time in cancer cells compared with normal cells.

In chapter 2, we explored the influences of four N^2-alkyl-dG lesions (alkyl = ethyl, *n*-butyl, isobutyl, or sec-butyl) on DNA replication in AB1157 *E. coli* cells and the isogenic strains with polymerases (Pol) II, IV, and V being individually or simultaneously knocked out. We found that N^2-Et-dG is slightly less blocking to DNA replication than the two isomeric N^2-*n*Bu-dG lesions, which display very similar replication bypass efficiencies. Additionally, Pol II and, to a lesser degree, Pol IV and Pol V are required for the efficient bypass of the N^2-alkyl-dG adducts, and none of these lesions were mutagenic.

In chapter 3, we used two independent approaches (proximity labeling and affinity pull-down) followed by LC-MS/MS analysis to profile Pol κ-interacting proteins at the proteome-wide level. We found that Pol κ interacts with DDX23 through the latter's transcription coactivator domain, and Pol κ ChIP-Seq analyses revealed that the polymerase is enriched at R-loop structure loci in chromatin. Moreover, Pol κ recruits DDX23 to R-loop sites in chromatin and promotes DDX23-mediated resolution of R-loops. We also observed augmented accumulation of R-loops in

Pol κ- and DDX23-deficient cells. Together, we discovered an interaction between Pol κ and DDX23, and revealed that the interaction facilitates resolution of R-loops in cells.

In chapter 4, we demonstrated N^2-alkyl-dG adducts induce R-loop structures accumulation on human chromatin by conducting IF and sp-KAS-seq analysis. Lesion-bearing vector-based R-ChIP-qPCR assay further illustrates R-loop structures formed on lesion insertion sites and CTAB assay results showed the lesion-induced R-loops strongly impede transcription process. And we also observed accumulation of DSBs after N^2-alkyl-dG incorporation in HEK293T and DDX23-deficient cells. Together, we conclude that N^2-alkyl-dG adducts trigger R-loop accumulation and further induce genome instability in human chromatin.

Since chemically damaged DNA has been linked to cancer and aging, future studies could be extended to map DNA damage formation and repair, thereby elucidating the distribution of damage on a genome-wide scale. Taking advantage of click chemistry, N^2-heptynyl-dG could be labeled with biotin-azide and enriched using streptavidin beads. This chemical labeling reaction was demonstrated to be highly specific and efficient regardless of flanking nucleotides and secondary structures. More importantly, taking advantage of blocking property of N^2-dG lesion toward DNA polymerase, a primer extension assay could be applied to identify a high-fidelity DNA polymerase that completely and precisely stalls just before the biotin-labeled N^2-dG, which may form the basis of detecting the exact position of N^2-dG.

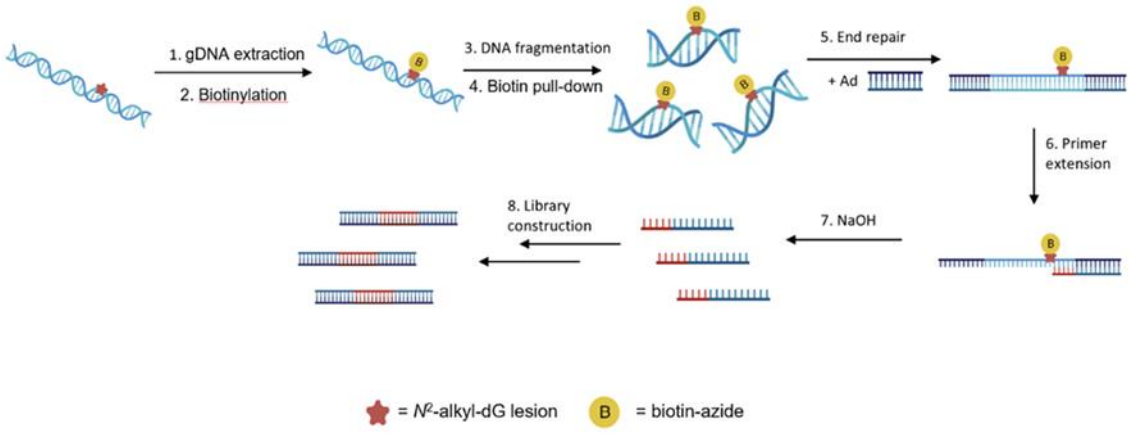

Figure 5-1. A schematic diagram of single-base resolution sequencing of N^2-heptynyl-dG in genomic DNA.

Future studies could also be extended to investigate the binding proteins of different DNA lesions. Employing DNA and protein cross-linking, DNA lesions and the corresponding binding proteins could be enriched based on their chemical properties. Furthermore, the occurrence of these binding proteins on DNA lesions would be of crucial importance to examine the biological significance of these proteins and their roles in cancer etiology and neuro-degenerative diseases. Further studies could also be extended to explore the other interacting partners of Pol κ and their functions. Besides DDX23, we found Pol κ also interacts with UBR5 and TRIM 21 in human cells, which are ubiquitin protein and E3 ligases, respectively. Additionally, the same proteomics approach can be used to study the interaction proteome of other TLS polymerases, such as Pol η and Rev1.