GENOME-WIDE EFFECTS OF DNA REPLICATION ON NUCLEOTIDE EXCISION REPAIR OF UV-INDUCED DNA LESIONS.

by
CEM AZGARI

Submitted to the Graduate School of Social Sciences
in partial fulfilment of
the requirements for the degree of Master of Arts

Sabancı University
August 2020

GENOME-WIDE EFFECTS OF DNA REPLICATION ON NUCLEOTIDE EXCISION REPAIR OF UV-INDUCED DNA LESIONS.

Approved by:

Dr. Öğr. Üyesi Ogün Adebali ....
(Thesis Supervisor)

Prof. Dr. Batu Erman ....

Prof. Dr. Halil Kavaklı ....

Date of Approval: August 28, 2020

ABSTRACT

GENOME-WIDE EFFECTS OF DNA REPLICATION ON NUCLEOTIDE EXCISION REPAIR OF UV-INDUCED DNA LESIONS.

CEM AZGARI

MOLECULAR BIOLOGY, GENETICS AND BIOENGINEERING M.A. THESIS, August 2020

Thesis Supervisor: Asst. Prof. Ogün Adebali

Keywords: Nucleotide excision repair, UV damage, (6-4)PP, CPD, XR-seq, Damage-seq, DNA replication, DNA strand asymmetry

Replication can cause unrepaired DNA damages to turn into mutations that might lead to cancer. Nucleotide excision repair is the leading repair mechanism that prevents melanoma cancers by removing UV-induced bulky adducts. However, the role of replication on nucleotide excision repair, in general, is yet to be clarified. Recently developed methods Damage-seq and XR-seq map damage formation and nucleotide excision repair events respectively, in various conditions. Here, we applied Damage-seq and XR-seq methods to UV-irradiated HeLa cells synchronized at two stages of the cell cycle: early S phase, and late S phase. We analyzed the damage and repair events along with replication origins and replication domains of HeLa cells. We found out that in both early and late S phase cells, early replication domains are more efficiently repaired relative to late replication domains. The results also revealed that repair efficiency favors the leading strand around replication origins. Moreover, we observed that the repair efficiency of the strands around origins is inversely correlated with the number of melanoma mutations.

ÖZET

DNA REPLİKASYONUNUN UV İLE İNDÜKLENEN DNA LEZYONLARININ NÜKLEOTİD EKSİZYONU ONARIMI ÜZERİNDEKİ GENOM ÇAPINDA ETKİLERİ.

CEM AZGARİ

MOLEKÜLER BİYOLOJİ, GENETİK VE BİYOMÜHENDİSLİK YÜKSEK LİSANS TEZİ, Ağustos 2020

Tez Danışmanı: Dr. Öğr. Üyesi Ogün Adebali

Anahtar Kelimeler: Nükleotid ekzisyon onarımı, UV hasarı, (6-4)PP, CPD, XR-seq, Damage-seq, DNA replikasyonu, DNA zinciri asimetrisi

Replikasyon, onarılmamış DNA hasarlarının kansere yol açabilecek mutasyonlara dönüşmesine neden olabilir. Nükleotid eksizyon onarımı, UV ile indüklenen hacimli DNA katımlarını ortadan kaldırarak melanom kanserlerini önleyen önde gelen onarım mekanizmasıdır. Ancak, replikasyonun nükleotid ekzisyon onarımındaki rolü henüz açığa kavuşturulmamıştır. Son zamanlarda geliştirilen yöntemler Damage-seq ve XR-seq sırasıyla, hasar oluşumu ve nükleotid eksizyon onarımı olaylarını çeşitli koşullar altında haritalandırabilmektedir. Burada, Damage-seq ve XR-seq yöntemlerini hücre döngüsünün erken ve geç S fazlarında senkronize edilip UV ile indüklenen HeLa hücrelerine uyguladık. HeLa hücrelerinin hasar ve onarım olaylarını replikasyon orijini ve replikasyon alanlarıyla birlikte analiz ettik. Hem erken hem de geç S fazlı hücrelerde, erken replikasyon alanlarının geç replikasyon alanlarına göre daha verimli bir şekilde onarıldığını bulduk. Sonuçlar ayrıca onarım verimliliğinin replikasyon orijinleri etrafında DNA'nın öncü ipliklerini desteklediğini ortaya koydu. Dahası, replikasyon orijini etrafındaki ipliklerin onarım etkinliğinin melanom mutasyonlarının sayısı ile ters orantılı olduğunu gözlemledik.

ACKNOWLEDGEMENTS

To my fiancée...

TABLE OF CONTENTS

LIST OF TABLES x

LIST OF FIGURES xi

1. INTRODUCTION 1

1.1. UV-Induced Damages in Humans 1

1.1.1. Cyclobutene Pyrimidine Dimers (CPDs) 2

1.1.2. Pyrimidine (6-4) Pyrimidone Photoproducts [(6-4)PPs] and their Dewar Valence Isomers 2

1.2. Nucleotide Excision Repair in Humans 3

1.2.1. Repair of UV-induced damages by Nucleotide Excision Repair 4

1.2.1.1. Damage Recognition 4

1.2.1.2. Dual Incision and Excision of Damaged Fragment ... 6

1.2.1.3. Re-synthesis and Ligation 6

1.2.2. NER associated diseases 7

1.3. Replication and its contribution to Mutagenesis 8

1.4. Mapping Damage Formation and Nucleotide Excision Repair Events using Damage-seq and Excision-seq (XR-seq) Methods, Respectively . 9

1.4.1. Damage-seq 9

1.4.2. Excision-seq (XR-seq) 10

2. The Scope of the Thesis 11

3. MATERIALS & METHODS 13

3.1. Materials 13

3.1.1. Samples 13

3.1.2. Programming Languages & Tools 13

3.1.3. Databases 13

3.2. Methods 13

3.2.1. Cell culture and treatments 16

3.2.2. Flow cytometry analysis 16

3.2.3. Damage-seq and XR-seq libraries preparation and sequencing . 16
3.2.4. Damage-seq sequence pre-analysis 17
3.2.5. XR-seq sequence pre-analysis 18
3.2.6. Dna-seq sequence pre-analysis 18
3.2.7. XR-seq and Damage-seq simulation 18
3.2.8. Quantification of melanoma mutations 19
3.2.9. Further analysis 19

4. RESULTS **20**
4.1. Genome-wide mapping of UV-induced damages and their repair synchronized at two stages of the cell cycle: early S phase, and late S phase 20
4.2. Early replication domains are repaired more efficiently than late replication domains, however, the repair rate of late replication domains elevates while replication proceeds. 22
4.3. Variety of chromatin states are associated with dominant repair. 23
4.4. ORIs display distinct melanoma mutation counts and strand asymmetry based on their replication domains. 25
4.5. Asymmetric damage around initiation zones causes asymmetric repair profiles. 28
4.6. Strand asymmetry of excision repair rate 29

5. DISCUSSION **32**
5.1. DNA replication elevates local nucleotide excision repair by mediating chromatin opening 33
5.2. UV-induced DNA damage, repair and mutagenesis display replicational strand asymmetry 35

BIBLIOGRAPHY **36**

6. APPENDIX **41**

LIST OF TABLES

LIST OF FIGURES

Figure 4.1. Experimental setup. 21
Figure 4.2. The shift of repair efficiency at replication domains during replication timing. 23
Figure 4.3. The effect of Chromatin States to repair efficiency of replication domains. 24
Figure 4.4. Tumor mutation profiles around replication origins and initiation zones for each replication domain. 27
Figure 4.5. Strand asymmetry around initiation zones caused by sequence content. 29
Figure 4.6. Repair rate asymmetry around initiation zones and replication domains. 31

Figure 5.1. Repair preferences of Nucleotide Excision Repair during replication. 34

Figure 6.1. Length distribution of excised oligomers of XR-seq samples. . . 41
Figure 6.2. Control figures of (6-4)PP asynchronized samples at 12 minutes. 42
Figure 6.3. Control figures of (6-4)PP early phased samples at 12 minutes. 42
Figure 6.4. Control figures of (6-4)PP late phased samples at 12 minutes. . 43
Figure 6.5. Control figures of CPD asynchronized samples at 12 minutes. . 43
Figure 6.6. Control figures of CPD early phased samples at 12 minutes. . . 44
Figure 6.7. Control figures of CPD late phased samples at 12 minutes. . . . 44
Figure 6.8. Control figures of CPD early phased samples at 120 minutes. . 45
Figure 6.9. Control figures of CPD late phased samples at 120 minutes. . . 45

LIST OF ABBREVIATIONS

CHR: Commission on Human Rights 2

DAW: Division for the Advancement of Women 5

DESA: Department of Economic and Social Affairs 9

1. INTRODUCTION

1.1 UV-Induced Damages in Humans

Ultraviolet (UV) light is the major cause of skin cancers in humans (Kiefer, 2007). It is a portion of the electromagnetic (EM) spectrum which is emitted from the sun together with visible light and heat. Based on its wavelength, UV light divides into three subgroups: UVA (wavelength 315-400 nm), UVB (wavelength 280-315 nm), and UVC (wavelength 100-280 nm). While the less energetic UVA makes up the majority of UV light passing the atmosphere, all UVC and approximately 90% of UVB is either blocked or absorbed by the ozone layer. Even in these conditions, we are not fully protected from the damaging effects of UV light. So that, UV irradiation accounts for approximately 30.000 DNA lesions' formation per cell per hour.

The most abundant UV lesions in cellular DNA are pyrimidine dimers (Kielbassa, Roza & Epe, 1997), which are formed by the covalent bonds between the adjacent pyrimidines (Whitmore, Potten, Chadwick, Strickland & Morison, 2001). Different in their chemical structure, two types of pyrimidine dimers are exist; one is called cyclobutene pyrimidine dimers (CPDs), and the other is called pyrimidine (6-4) pyrimidone photoproducts [(6-4)PPs]. While both UVC and UVB can induce these dimer formations, UVA is only capable of induce CPDs. Nonetheless, UVA induction can convert already formed (6-4)PPs into their Dewar valence isomers. Moreover, UVA can induce oxidative DNA damages through photosensitized reactions (Hu & Adar, 2017). Thanks to the development of time resolved spectroscopy techniques in recent years, dynamics of pyrimidine dimer formation is well known. The formation and biological properties of UV lesions will be briefly discussed in the subsections below.

1.1.1 Cyclobutene Pyrimidine Dimers (CPDs)

CPDs are the most frequent pyrimidine dimers that are arising from the covalent linkages between the consecutive pyrimidines, and it is characterized by the four-member ring structure that are double bonded from the pyrimidine 5 and 6 (Whitmore et al., 2001). In vivo, CPDs can be observed in four different configurations: cis-syn, cis-anti, trans-syn, or trans-anti. (Khattak & Wang, 1972) While it is generally observed in cis-syn form when the DNA is double-stranded (Wacker, Dellweg, Träger, Kornhauser, Lodemann, Türck, Selzer, Chandra & Ishimoto, 1964), in denatured DNA and single-stranded regions, trans-syn configuration exists (Taylor & Brockie, 1988). Although it is rare, nonadjacent pyrimidines can also form CPDs in single-stranded regions (Nguyen & Minton, 1988). Moreover, different configurations can affect the ability of repair enzymes to recognize these lesions and correct them, which cause mutability differences between the configurations (Friedberg, Walker, Siede & Wood, 2005).

Apart from the configuration of the lesions, their dipyrimidine doublets (TT, TC, CT, and CC) can contribute to CPD formation at different rates depending on the type of UV exposure or the nucleotide content of the DNA. According to the study of Douki and Cadet, under UVC and UVB exposure, double-stranded mammalian DNA produces TT, TC, CT, and CC CPDs in 100:50:25:10 ratios, respectively (Douki & Cadet, 2001). While TT CPDs accounts for more than half of the total CPDs after the exposure of UVC and UVB, for UVA exposure, this ratio rises to 90% (Mouret, Philippe, Gracia-Chantegrel, Banyasz, Karpati, Markovitsi & Douki, 2010). On the other hand, TT CPDs are the abundant products of UV exposure for mammalian DNA, but the abundance might be greatly influenced by the GC percentage of the DNA. For example, in the bacterial DNA that possess a rich GC percentage, TT CPDs are the minor products of UV exposure (Patrick, 1977).

1.1.2 Pyrimidine (6-4) Pyrimidone Photoproducts [(6-4)PPs] and their Dewar Valence Isomers

(6-4)PPs form with occurrence of a pyrimidone ring by the bonding between C6 position of the 5'-end base and C4 position of the 3'-end base. In fact, this structure forms indirectly following the UV exposure, after a cyclic reaction intermediate, which can be either an oxetane if thymine is the 3'-end base, or azetidine if cytosine is the 3'-end base. Because of its indirect formation, (6-4)PPs appear thousand times slower than CPDs (Schreier, Schrader, Koller, Gilch, Crespo-Hernández, Swaminathan, Carell, Zinth & Kohler, 2007).

Under UVC and UVB exposure, formation of (6-4)PPs is approximately five time less than that of CPDs (Douki & Cadet, 2001). Moreover, TT dipyrimidines that are the most abundant sites for CPDs are less frequent for (6-4)PPs. Instead, TC and CCs are the frequent sites for (6-4)PPs, while CT (6-4)PPs are uncommon. Another unique property of (6-4)PPs is its conversion into Dewar valence isomers with the photoisomerization process (Taylor & Cohrs, 1987). Although UVB irradiation can trigger the process, with the combination of UVB and UVA exposure, the yield increases significantly.

1.2 Nucleotide Excision Repair in Humans

Throughout the generations, cells manage to evolve highly specialized repair mechanisms to cope with a variety of lesions that threaten the genome integrity and survival. Considering the diversity of these lesions, it would be unexpected to have only a single mechanism that can preserve the integrity of the genome. Hence, there are several repair mechanisms that cells utilize which are eminently conserved between species. Due to the removal of both strands, repair of a double strand break is demanding. There are two mechanisms that can be triggered by double strand breaks: homologous recombination, and non-homologous end-joining. Homologous recombination uses the sister-chromatid as a template to repair double strand breaks in an error-free manner. In addition, if sister-chromatid is not available for use, non-homologous end-joining directs the fusion of broken ends in an error-prone manner. Although being error-prone, non-homologous end-joining is the dominant mechanism for double strand break repair in mammals. Reasons of this dominancy are the distant proximity of chromatids to each other, and the DNA folding that makes the homologous sequence less reachable. In addition, imperfect matches by homologous recombination can lead to tragic outcomes such as creating repeated sequences (Li, Wehrenberg, Waldman & Waldman, 2018).

On the other hand, when a damage occurs at a single strand, the opposite strand can be used as a template. In such cases, DNA excision repair mechanisms remove the lesion site and re-synthesize the gap using the template strand. While oxidation, deamination and alkylation damages are repaired by base excision repair (Klungland, Höss, Gunz, Constantinou, Clarkson, Doetsch, Bolton, Wood & Lindahl, 1999), mismatches that escape proofreading are identified and corrected by mismatch repair (Modrich, 1997). And lastly, bulky adducts caused by UV irradiation, environmental mutagens, and chemotherapeutic agents are removed by nucleotide excision repair (Reardon & Sancar, 2005). Nucleotide excision repair contains two sub pathways that differ from each other at the damage recognition step: Global Repair (GR) and Transcription-Coupled Repair (TCR). TCR is specialized in recognizing adducts in transcribed regions, while GR can recognize bulky adducts at any site. Subsections below will address the assembly and main properties of nucleotide excision repair in more detail.

1.2.1 Repair of UV-induced damages by Nucleotide Excision Repair

Identified firstly at E. coli by two independent studies published in 1964 (Boyce & Howard-Flanders, 1964; Setlow & Carrier, 1964), nucleotide excision repair can repair variety of bulky adducts from UV-induced pyrimidine dimers to chemotherapeutic agents such as cisplatin (Yimit, Adebali, Sancar & Jiang, 2019). Although repair mechanisms are highly conserved among the species, nucleotide excision repair in humans appeared to be surprisingly different from that of E. coli. While E. coli contains three proteins (UvrA, B, C) for the incision of damaged fragments, human nucleotide excision repair has sixteen proteins for the task. More interestingly, there is not an evolutionarily relevance between these human and E. coli proteins. In addition, the excised fragments are usually around 12 nucleotides long in E. coli. For humans, the length of these fragments are around 30 nucleotides (Sancar, 2016). Human nucleotide excision repair can be generally discussed in three steps: 1) damage recognition, 2) dual incision and excision of damaged fragments, and 3) re-synthesis and ligation.

1.2.1.1 Damage Recognition

As it is mentioned earlier, GR and TCR have distinct damage recognition steps. GR scans the whole genome to detect helix distortions caused by bulky adducts, whereas TCR responds only to a stalled RNA polymerase II during transcription.

In GR, three proteins (XPC, RAD23B, CETN2) work in coordination to recognize the lesion site (Sugasawa, Ng, Masutani, Iwai, van der Spek, Eker, Hanaoka, Bootsma & Hoeijmakers, 1998). XPC is the first protein to interact with the lesion by binding to the small single-stranded DNA (ssDNA) that is left unpaired due to the pyrimidine dimer formation at the opposite strand. The ability of XPC to bind unpaired ssDNA enables GR to detect a variety of lesions, since the unpaired ssDNA is a common characteristic of bulky adducts. After XPC binding, RAD23B and CETN2 interact with and stabilize XPC. However, helix distortions must be apparent to XPC for an efficient detection. (6-4)PPs are recognized relatively in ease because of having a prominent distortion (Mizukoshi, Kodama, Fujiwara, Furuno, Nakanishi & Iwai, 2001), whereas the distortion of CPDs cause only a 9^{o} unwinding with a 30^{o} bent (Park, Zhang, Ren, Nadji, Sinha, Taylor & Kang, 2002), which can be considered mild. For the detection of CPDs, proteins DDB1 and DDB2 form a complex called ultraviolet radiation–DNA damage-binding protein (UV-DDB). The complex directly interacts with the lesion, and DDB2 kinks the lesion to increase unwinding (Scrima, Koníčková, Czyzewski, Kawasaki, Jeffrey, Groisman, Nakatani, Iwai, Pavletich & Thomä, 2008), as a result the ssDNA becomes detectable for XPC.

The recognition mechanism of TCR is triggered by the blockage of RNA polymerase II (RNAPII), which transcribes the active gene during transcription elongation. When RNAPII stalls following an encounter with a lesion, it subsequently recruits the nucleotide excision repair proteins (Svejstrup, 2002). Afterwards, RNAPII dynamically interacts with UV-stimulated scaffold protein A (UVSSA), ubiquitin-specific-processing protease 7 (USP7), Cockayne syndrome protein CSB. CSB is an ATP-dependent chromatin remodeling factor that contains a helicase motif, surprisingly without a helicase activity (Selby & Sancar, 1997b). Moreover, studies in early 2000s revealed that point mutations in the ATPase domain of CSB protein significantly cripples the cell's ability to escape the inhibited RNA synthesis (Citterio, Rademakers, van der Horst, van Gool, Hoeijmakers & Vermeulen, 1998; Muftuoglu, Selzer, Tuo, Brosh Jr & Bohr, 2002), which suggests that CSB plays a key role for the TCR assembly. Furthermore, recruitment of repair factors that work on incision of the damaged fragment also mediated by CSB (Fousteri, Vermeulen, van Zeeland & Mullenders, 2006). More identified functions of CSB include transcription elongation, chromatin maintenance and remodeling, histone tail binding, and strand annealing (Selby & Sancar, 1997a). Another important Cockayne syndrome protein is CSA, which is also recruited by CSB. CSA mediates the recruitment of PCNA, RFC and pol δ. Therefore, it is a key protein for the later events of the repair.

The recruited core nucleotide excision repair factors and some TCR specific factors such as UV-stimulated scaffold protein A (UVSSA), ubiquitin-specific-processing protease 7 (USP7), XPA-binding protein 2 (XAB2) and high mobility group nucleosome-binding domain-containing protein 1 (HMGN1), gather on the lesion site where RNAPIIo stalls. However, because RNAPIIo stalls on the lesion, it covers the lesion so that the TCR complex cannot reach it (Tornaletti, Reines & Hanawalt, 1999). To proceed, RNAPIIo should somehow move from the 35 nucleotides length of strand where it is positioned. There are three proposed mechanisms for that purpose which are degradation, dissociation and backtracking. Because backtracking is already known to be occurring at transcription proofreading and at natural transcription pause sites, it is the most accepted mechanism among these three (Marteijn, Lans, Vermeulen & Hoeijmakers, 2014).

1.2.1.2 Dual Incision and Excision of Damaged Fragment

After RNAPIIo backtracks, transcription initiation factor IIH (TFIIH) initiates to unwind DNA with its helicase subunits. The TFIIH complex is formed of 10 proteins. While XPB and XPD have helicase activity, CDK-activating kinase (CAK) subcomplex is responsible for the initiation of TFIIH complex. The initiation is also known as DNA damage verification step which is the last reversible step of nucleotide excision repair (Marteijn et al., 2014). With the initiation of TFIIH complex, the lesion becomes ready to be removed. Then XPF-ERCC1 and XPC endonucleases interact with the lesion site to catalyze the lesion from two sides together with the TFIIH complex. Meanwhile, replication protein A (RPA) not only protects the non-damaged single strand, but also interacts with and coordinates most subunits of TFIIH complex. The cleavage of the lesion site that yields 22-30 nucleotide long single stranded gap, is termed dual incision (Marteijn et al., 2014).

1.2.1.3 Re-synthesis and Ligation

After the dual incision, the occurred gap must be filled with the ligation process. During replication, the proteins proliferating cell nuclear antigen (PCNA), replication factor C (RFC), DNA pol δ, DNA pol ϵ and DNA ligase 1 mediates re-synthesis and ligation. However, if the cell is non-replicating, then DNA pol κ and XRCC1–DNA ligase 3 fill the gap (Marteijn et al., 2014).

1.2.2 NER associated diseases

There are three human diseases that are known to be directly associated with nucleotide excision repair. These diseases are xeroderma pigmentosum (XP), cockayne syndrome (CS) and trichothiodystrophy (TTD) (De Boer & Hoeijmakers, 2000; Lehmann, 2003). XP discovered in 1968 as a hereditary disease that causes a defective nucleotide excision repair (Cleaver, 1968). XP patients are extremely photosensitive, so that they have approximately 5000-fold increased risk of UV-induced skin cancer. Dry parchment skin and pigmentation related anomalies are some of the hallmarks of this disorder (De Boer & Hoeijmakers, 2000). Seven genes that are associated with the disease, known as XP complementation groups (XP-A, B, C, D, E, F, G) (Cleaver & Bootsma, 1975), and proteins that are produced by all these genes have a role in GR. Except XPC and XPE, they are also involved in TCR (Van Hoffen, Venema, Meschini, Van Zeeland & Mullenders, 1995).

CS first reported in 1936 as a disease related to deafness and dwarfism (Cockayne, 1936). In the upcoming years, problems at joints, vision, and calcifications in the brain are further reported (Cockayne, 1946; Neill & Dingwall, 1950). Moreover, these patients have aging related issues, and like XP patients, they are photosensitive, though not as severe as XP patients, therefore, their risk of having UV-induced skin cancer is not increased. As a consequence of all these abnormalities, most severe types of CS patients have a lifespan of as short as 7 years. Two genes, CSA and CSB are known to be related to the disease, which are both TCR proteins. Thereby, it was thought that CS patients are TCR defective. However, since TCR deficiency is not enough to explain all these severe symptoms alone, a deficiency in transcription is also argued (Drapkin, Reardon, Ansari, Huang, Zawel, Ahn, Sancar & Reinberg, 1994).

TTD patients can display a broad range of symptoms from having brittle hair to low fertility and impaired intelligence. If the disorder is caused by one of the XPB, XPD or TTDA genes, which are all code for a component of TFIIH complex, TTD patients can become nucleotide excision repair deficient, hence photosensitive. Even though TFIIH complex can be functional, the levels of TFIIH complex decreases significantly (Giglia-Mari, Miquel, Theil, Mari, Hoogstraten, Ng, Dinant, Hoeijmakers & Vermeulen, 2006).

1.3 Replication and its contribution to Mutagenesis

Owing to many potential origins of replication (ORIs), a mammalian cell replicates in approximately 10 hours (Takebayashi, Ogata & Okumura, 2017). During the cell division, only a portion of these ORIs fires, and they fire in an asynchronized manner except the ORIs that are in proximity to each other. By firing simultaneously, these close packed ORIs coordinates the replication of regions longer than mega bases, termed as "replication domains" (Jackson & Pombo, 1998). Replication domains are divided into 4: early replication domains, late replication domains and the zones between these domains are up transition zones and down transition zones (Farkash-Amar, Lipson, Polten, Goren, Helmstetter, Yakhini & Simon, 2008; Hansen, Thomas, Sandstrom, Canfield, Thurman, Weaver, Dorschner, Gartler & Stamatoyannopoulos, 2010; Hiratani, Ryba, Itoh, Yokochi, Schwaiger, Chang, Lyou, Townes, Schübeler & Gilbert, 2008; Koren, Handsaker, Kamitaki, Karlić, Ghosh, Polak, Eggan & McCarroll, 2014; Nakayasu & Berezney, 1989; O'keefe, Henderson & Spector, 1992). Generally, the interior regions of the nucleus are replicated earlier than nuclear periphery regions, thus located at early replication domains (Dimitrova & Berezney, 2002). Multiple studies indicated that these domains are differ each other in the mutation frequencies. Suggested by the genome-wide analysis of mutation rates, early replication domains have reduced levels of mutation comparing to late replication domains (Lawrence, Stojanov, Polak, Kryukov, Cibulskis, Sivachenko, Carter, Stewart, Mermel, Roberts & others, 2013; Stamatoyannopoulos, Adzhubei, Thurman, Kryukov, Mirkin & Sunyaev, 2009). Also, in most cancers, base substitution mutation elevates in late replication domains (Schuster-Böckler & Lehner, 2012).

Replication is driven by replication forks which are formed when a predefined ORI fires (Langston, Indiani & O'Donnell, 2009). Replication fork usually proceeds bidirectionally, with the coordinated work of polymerases ϵ and δ. During the movement of the fork, polymerases ϵ continuously synthesizes leading strands, whereas polymerase δ discontinuously synthesizes lagging strands. Moreover, bidirectionality creates an asymmetric work labor, so that two polymerases work on opposite strands towards different directions. In other words, in the left replicating fork, polymerases ϵ proceeds on the plus strand, while in the right replicating fork, it progresses on the minus strand. Studies suggests that this asymmetric progress of polymerases around the associated ORI are reflected to the mutation profiles, where lagging strand reported to harbor more mutations than leading strand (Haradhvala, Polak, Stojanov, Covington, Shinbrot, Hess, Rheinbay, Kim, Maruvka, Braunstein & others, 2016; Lujan, Williams, Pursell, Abdulovic-Cui, Clark, McElhinny & Kunkel, 2012; Reijns, Kemp, Ding, de Procé, Jackson & Taylor, 2015; Shinbrot, Henninger, Weinhold, Covington, Göksenin, Schultz, Chao, Doddapaneni, Muzny, Gibbs & others, 2014).

The occurred asymmetry on mutation profiles is reasoned by the error-prone bypass mechanism on the leading strand that makes it vulnerable to mutations (Seplyarskiy, Akkuratov, Akkuratova, Andrianova, Nikolaev, Bazykin, Adameyko & Sunyaev, 2019). Other studies argued that the attachment of helicase to leading strand increases the damage response, thus leading to effective repair of the strand (Hedglin & Benkovic, 2017; Yeeles, Poli, Marians & Pasero, 2013). Furthermore, many mutational signatures are reported to have a significant replication strand asymmetry (Tomkova, Tomek, Kriaucionis & Schuster-Böckler, 2018).

1.4 Mapping Damage Formation and Nucleotide Excision Repair Events using Damage-seq and Excision-seq (XR-seq) Methods, Respectively

Mapping of UV-induced damages and their repair is assential to understand the role of nucleotide excision repair on mutagenesis. Since the birth of the field of DNA repair, which began with the discovery of photolyase in 1958 (Rupert, Goodgal & Herriott, 1958; Sancar, 2016), many methods are introduced to map DNA damage and repair (Li & Sancar, 2020). However, not until the emergence of next-generation sequencing techniques, genome-wide mapping of DNA damage and repair at single-nucleotide resolution could be performed. Today, there are several methods that can accomplish the task. Among these methods, Damage-seq and Excision-seq (XR-seq) can map UV-induced DNA damages and repair of these damages by nucleotide excision repair, respectively, which will be explained in the subsections below.

1.4.1 Damage-seq

Damage-seq mechanism can sensitively detect a variety of DNA lesions such as CPDs, (6-4)PPs, and cisplatins, mainly using the DNA polymerase II stalling to its advantage (Hu, Lieb, Sancar & Adar, 2016). In fact, the method can be adapted to any DNA damage that stalls DNA polymerase II, where the damage-specific antibody is present (Sancar, 2016). After the induction of the damage, the genomic DNA is sonicated, ligated to first primers, and denaturated. Then, damage sites are immunoprecipitated by damage-specific antibodies and enriched. Following the enrichment, a biotinylated primer is annealed and extended by a polymerase called

Q5 DNA polymerase, which extends the primer until it reachs the damage without synthesizing the site of the damage. Next, a second adopter is ligated to the extended primer for amplification by PCR. Lastly, the amplified oligomers can be sequenced and analyzed.

1.4.2 Excision-seq (XR-seq)

XR-seq method can measure the repair of DNA damages that is coordinated by nucleotide excision repair, using the 22-30 nt long exiced oligomers that are produced after the dual incision of lesion site (Hu, Li, Adebali, Yang, Oztas, Selby & Sancar, 2019; Hu et al., 2016). Excised oligomers are immunoprecipitated by TFIIH and ligated by adaptors from both sides. Next, the oligomers are filtered according to the damage of interest by immunoprecipitating with damage-specific antibodies. Then, using photolyases, lesions of the left oligomers are reversed for a proper PCR amplification process and the oligomers are sequenced.

2. The Scope of the Thesis

Nucleotide excision repair is the sole mechanism for the removal of bulky adducts. In this study, to assess the influence of replication along with nucleotide excision repair on mutation distribution across replicated sites, we analyzed the Damage-seq and eXcision Repair sequencing (XR-seq) data. Damage and repair maps are generated for cyclobutane pyrimidine dimers (CPDs) and pyrimidine-pyrimidone (6-4) photoproducts [(6-4)PPs] from UV-irradiated HeLa cells synchronized at two stages of the cell cycle: early S phase, and late S phase. Damage-seq locates and quantifies the regions of UV induced CPD and (6-4)PP damages, while XR-seq captures excised oligomers of the damage site that are removed by the nucleotide excision repair. Two methods combined provide the genome-wide distribution of UV-induced damages and the differential repair frequency of these damage sites.

Initially, we examined the quality of the reads that are produced by Damage-seq and XR-seq methods. After quality filtering and performing pre-analysis of damage and repair reads, we prepared bed files containing the positions of damage and repair events on the human genome. Then, we used these bed files together with data sets obtained from public sources to compare the repair rate of nucleotide excision repair at different regions.

In the first part of the study, we mapped the damage and repair events to the replication domains, where closely packed origin of replications fire in a synchronized manner, resulting in simultaneous replication of these Mb-sized regions. Then, we normalized repair events with corresponding damage quantities to eliminate the potential bias caused by the damage formation. By doing so, we managed to observe the differential repair rate between replication domains at different time points on a wide scale. We performed a similar analysis using chromatin states of HeLa cells and examined how chromatin states effect the repair rate of replication domains, while moving early to late S phase of cell cycle.

Secondly, we aim to find whether nucleotide exicision repair contribute to a replicative strand asymmetry. Because nucleotide excision repair is highly associated with melanoma cancers, replicative strand asymmetry of nucleotide excision repair can correlate with the mutation profiles of melanoma cancers. We retrieved a somatic melanoma mutations data, and quantified the mutations on approximately 20 kb-sized initiation zones where origin of replications closely positioned. We further separated these initiation zones into their corresponding replication domains before quantifying the mutations. This method enabled us both to compare the mutation count differences of replication domains, and to observe the mutational strand asymmetry on initiation zones. Next, we examined the strand asymmetry of damage and repair events separately on initiation zones. To see if nucleotide composition of initiation zones is contributing to the strand asymmetry, we simulated Damage-seq and XR-seq reads, and compared the signal levels of these reads on initiation zones as well. Lastly, we calculated the repair rate by normalizing repair events with damage quantities and mapped them to observe the asymmetry of repair rate.

3. MATERIALS & METHODS

3.1 Materials

3.1.1 Samples

3.1.2 Programming Languages & Tools

3.1.3 Databases

3.2 Methods

The experiments were performed at SancarLab by our collaborators, whereas analyses of the data were carried by us.

cell line	**product**	**method**	**release**	**time**	**replicate**
HeLa-S3	CPD	XR-seq	early	120	A
HeLa-S3	CPD	XR-seq	late	120	A
HeLa-S3	CPD	XR-seq	early	120	B
HeLa-S3	CPD	XR-seq	late	120	B
HeLa-S3	CPD	Damage-seq	early	120	A
HeLa-S3	CPD	Damage-seq	late	120	A
HeLa-S3	CPD	Damage-seq	early	120	B
HeLa-S3	CPD	Damage-seq	late	120	B
HeLa-S3	(6-4)PP	XR-seq	async	12	A
HeLa-S3	(6-4)PP	XR-seq	async	12	B
HeLa-S3	(6-4)PP	XR-seq	early	12	A
HeLa-S3	(6-4)PP	XR-seq	early	12	B
HeLa-S3	(6-4)PP	XR-seq	late	12	A
HeLa-S3	(6-4)PP	XR-seq	late	12	B
HeLa-S3	CPD	XR-seq	async	12	A
HeLa-S3	CPD	XR-seq	async	12	B
HeLa-S3	CPD	XR-seq	early	12	A
HeLa-S3	CPD	XR-seq	early	12	B
HeLa-S3	CPD	XR-seq	late	12	A
HeLa-S3	CPD	XR-seq	late	12	B
HeLa-S3	(6-4)PP	Damage-seq	async	12	A
HeLa-S3	(6-4)PP	Damage-seq	async	12	B
HeLa-S3	(6-4)PP	Damage-seq	early	12	A
HeLa-S3	(6-4)PP	Damage-seq	early	12	B
HeLa-S3	(6-4)PP	Damage-seq	late	12	A
HeLa-S3	(6-4)PP	Damage-seq	late	12	B
HeLa-S3	CPD	Damage-seq	async	12	A
HeLa-S3	CPD	Damage-seq	async	12	B
HeLa-S3	CPD	Damage-seq	early	12	A
HeLa-S3	CPD	Damage-seq	early	12	B
HeLa-S3	CPD	Damage-seq	late	12	A
HeLa-S3	CPD	Damage-seq	late	12	B

Programming Languages and Tools	**Description**	
Bash	a shell compatible command language	
Python	a high-level, general purpose programming language	
R	a language and an environment for graphics and statistics	
Cutadapt	detects and cuts adaptor sequences	
Bowtie2	a fast and memory-efficient sequence aligner	
Samtools	a suit that contains utilities to interact with and manipulate high-throughput sequencing data	
Bedtools	a set of utilities to perform genomic analysis	co
BedGraphToBigWig	converts bedGraph files to bigWig	cre
Art	a simulation tool that creates a synthetic high-throughput sequencing data	simu

Databases	**Data Obtained**	**Source**
The European Bioinformatics Institute FTP Server	Genome Reference Consortium Human Build 37 (GRCh37)	ref
Gene Expression Omnibus (GEO)	Repli-Seq data of HeLa-S3 (accession no: GSE53984), SNS-Seq data of HeLa-S3 (accession no: GSE37757)	ref
UCSC Genome Browser	ChromHMM segmentation from HeLa-S3 ChIP-Seq data	ref
Sequence Read Archive (SRA)	OK-Seq data (accession no: SRP065949)	ref
International Cancer Genome Consortium (ICGC)	Simple somatic mutations of Melanoma	ref

3.2.1 Cell culture and treatments

HeLa-S3 cell lines that were purchased from ATCC were cultured in DMEM medium supplemented with 10% FBS and 1% penicillin/streptomycin at 37°C in a 5% atmosphere CO_2 humidified chamber. By double-thymidine treatment, cells were synchronized at late G1 phase, and released into S phase after the removal of thymidine. Thymidine at 50% confluence was added to the cells to a final concentration of 2 mM for the initial thymidine treatment. After 18 hours, the cells were washed with PBS for their release 18 hours after the initial thymidine treatment, and cultured in fresh medium for 9 hours. Then for 15 hours, cells were treated with 2mM thymidine and released into S phase for designated time before UV irradiation. Cells were irradiated with $20J/m^2$ of UVC, then collected either immediately or after incubation at 37°C for designated time for the following assays.

3.2.2 Flow cytometry analysis

HeLa-S3 cell lines were initially trypsinized, and then PBS washed. After washing, for 2 hours, cells were fixed in 70% (v/v) ethanol at -20°C, then for 30 minutes, stained in the staining solution at room temperature. Lastly, the progression of the cells throughout the S phase was analyzed by a flow cytometer.

3.2.3 Damage-seq and XR-seq libraries preparation and sequencing

After HeLa-S3 cell lines were harvested in ice-cold PBS at designated time, Damage-seq and XR-seq methods were applied. For Damage-seq, using PureLink Genomic DNA Mini Kit, genomic DNA was taken out and then, cut into fragments by sonication using Q800 Sonicator. After sonication, DNA fragments ($1\mu g$) were subjected to end repair, dA-tailing and ligation using the first adaptor. Then, the fragments were denaturated and immunoprecipitated with either anti-(6-4)PP or anti-CPD antibody. A primer called Bio3U was bound to the fragment and extended with Q5 DNA polymerase until the primer reaches the lesion site. Next, the extended primer fragments were purified and annealed to oligo SH for subtractive hybridization process. After the substractive hybridization, oligo SH was removed using streptavidin C1 and the fragments were ligated to the second adapter for PCR amplification process. For XR-seq, cells were lysed with a homogenizer and centrifuged to remove chromatin DNA. To extract the nucleotide excision repair products, lysed cells were immunoprecipitated with anti-XPG antibody, which precipitates the excision products. Then, purified fragments were ligated with adaptors from both ends. The fragments were further immunoprecipitation with either anti-(6-4)PP or anti-CPD antibody and lesion sites were repaired by photolyase. After PCR amplification and gel purification, the products were sequenced via Hiseq 2000/2500 platform by the University of North Carolina High-Throughput Sequencing Facility, or Hiseq X platform by the WuXiNextCODE Company.

3.2.4 Damage-seq sequence pre-analysis

The sequenced reads with adapter sequence GACTGGTTCCAATTGAAAGT-GCTCTTCCGATCT at 5' end, were discarded via cutadapt with default parameters for both single-end and paired-end reads (Martin, 2011). The remaining reads were aligned to the hg19 human genome using bowtie2 with 4 threads (-p) (Langmead & Salzberg, 2012). For paired-end reads, maximum fragment length (-X), which means the maximum accepted total length of mated reads and the gap between them, was chosen as 1000. Using samtools, aligned paired-end reads were converted to bam format, sorted using samtools sort -n command, and properly mapped reads with a mapping quality greater than 20 were filtered using the command samtools view -q 20 -bf 0x2 in the respective order (Li et al.). Then, resulting bam files were converted into bed format using bedtools bamtobed -bedpe -mate1 command (Quinlan & Hall, 2010). The aligned single-end reads were directly converted into bam format after the removal of low quality reads (mapping quality smaller than 20) and further converted into bed format with bedtools bamtobed command (Quinlan & Hall, 2010). Because the exact damage sites should be positioned at two nucleotides upstream of the reads (Li et al.), bedtools flank and slop command were used to obtain 10 nucleotide long positions bearing damage sites at the center (5. and 6. positions) (Quinlan & Hall, 2010). The reads that have the same starting and ending positions, were reduced to a single read for deduplication and remaining reads were sorted with the command sort -u -k1,1 -k2,2n -k3,3n. Then, reads that did not contain dipyrimidines (TT, TC, CT, CC) at their damage site (5. and 6. positions) were filtered out to eliminate all the reads that do not harbor a UV damage. Lastly, only the reads that were aligned to common chromosomes (chromosome 1-22 + X) were held for further analysis.

3.2.5 XR-seq sequence pre-analysis

The adaptor sequence TGGAATTCTCGGGTGCCAAG-GAACTCCAGTNNNNNNACGATCTCGTATGCCGTCTTCTGCTTG at the 3' of the reads were trimmed and sequences without the adaptor sequences were discarded using cutadapt with default parameters (Martin, 2011). Bowtie2 was used with 4 threads (-p) to align the reads to the hg19 human genome (Langmead & Salzberg, 2012). Then reads with mapping quality smaller than 20 were removed by samtools (Li et al.). Bam files obtained from samtools were converted into bed format by bedtools (Quinlan & Hall, 2010). Multiple reads that were aligned to the same position, were reduced to a single read to prevent duplication effect and remaining reads were sorted with the command sort -u -k1,1 -k2,2n -k3,3n. Lastly, only the reads that were aligned to common chromosomes were held for further analysis.

3.2.6 Dna-seq sequence pre-analysis

Paired-end reads were aligned to hg19 human genome via bowtie2 with 4 threads (-p) and maximum fragment length (-X) chosen as 1000 (Langmead & Salzberg, 2012). Sam files were converted into bed format as it was performed at damage-seq paired-end reads. Duplicates were removed and reads were sorted with sort -u -k1,1 -k2,2n -k3,3n command. Lastly, the reads that did not align to the common chromosomes were discarded.

3.2.7 XR-seq and Damage-seq simulation

Art simulator was used to produce synthetic reads with the parameters -l 26 -f 2, -l 10 -f 2 for XR-seq and Damage-seq, respectively (Huang, Li, Myers, & Marth). To better represent our filtered real reads, read length (-l) parameter was chosen as the most frequent read length after pre-analysis done. The .fastq file that Art produced, were filtered according to our reads by calculating a score using nucleotide frequency of the real reads and obtaining most similar 10 million simulated reads. The filtering was done by filter_syn_fasta.go script, which is available at the repository: https://github.com/compGenomeLab/lemurRepair. Filtered files were preceded by pre-analysis again for further analysis.

3.2.8 Quantification of melanoma mutations

Melanoma somatic mutations of 183 tumor samples were obtained from the data portal of International Cancer Genome Consortium (ICGC) as compressed .tsv files which is publicly available at https://dcc.icgc.org/releases/release_28/Projects/MELA-AU. Single base substitution mutations were extracted, and only the mutations of common chromosomes were used. To obtain the mutations that could be caused by UV-induced photoproducts, C -> T mutations that have a pyrimidine at the upstream nucleotide was further extracted. Later on, mutations were quantified on 20 kb long initiation zones that were separated into their corresponding replication domains using bedtools intersect command with the -wa -c -F 0.5 options.

3.2.9 Further analysis

In order to separate a region data (replication domains, initiation zones, or replication origins) into chosen number of (201) bins, the start and end positions of all the regions set to a desired range with the unix command: Then, any intersecting regions or regions crossing the borders of its chromosomes were filtered to eliminate the possibility of signal's canceling out effect. After that, bedtools makewindows command was used with the -n 201 -i srcwinnum options to create a .bed file containing the bins. To quantify the XR-seq and Damage-seq profiles on the prepared .bed file, bedtools intersect command was used to intersect as it was performed for mutation data. Then all bins were aggregated given their bin numbers, and the mean of the total value of each bin were calculated. Lastly RPKM normalization was performed and the plots were produced using ggplot2 in R programming language.

4. RESULTS

4.1 Genome-wide mapping of UV-induced damages and their repair

synchronized at two stages of the cell cycle: early S phase, and late S phase

This paper presents an experimental setup followed by bioinformatic analysis of genomic data, where we purified and sequenced fragments of UV-induced damages and their repair in HeLa cells that are synchronized either at early S phase or at late S phase. After synchronizing cells using double-thymidine treatment, we further treated cells with 20J/m2 UV-B exposure. Immediately after the exposure, we adopted Damage-seq to quantify occurred damages by the exposure, before nucleotide excision repair initiates. To quantify repair, we adopted XR-seq and quantified CPD repair at 12 minutes and 2 hours; while (6-4)PP repair were quantified only at 12 minutes (Figure 1A). Finally, we performed each experiment twice to obtain biological replicates.

Quality control analyses were performed on early S phased (6-4)PPs at 12 minutes (Figure 1B-D) and other samples (Figure S2-8) indicates high data qualities and consistent results between replicates. In agreement with the dual incision mechanism of nucleotide excision repair (J. C. Huang, Svoboda, Reardon, & Sancar, 1992; Li et al., 2017; Reardon & Sancar, 2005), XR-seq oligomers are in the size range of 20-30 nt, with a median of 26 nt. Moreover, dipyrimidine content of 26 nt oligomers enriches at position 19-20 (Figure 1B) where the DNA lesion occurs (J. C. Huang et al., 1992). Also, (6-4)PP samples exhibit high levels of TC dipyrimidine repair (Figure 1B, S2-4) whereas CPD samples exhibit elevated TT dipyrimidine repair (Figure S5-9), which are the most abundant sites for formation of these photoproducts (Mouret et al., 2010). Because this study focuses on the global repair, contribution of transcription-coupled repair can create a bias. Importantly, the repair levels at transcribed and non-transcribed strand are equivalent (Figure 1C, S2-9), suggesting little or no contribution of transcription-coupled repair. Correlation plots between the biological replicates indicates reasonable reproducibility, having correlation coefficients 0.86 and above (Figure 1D, S2-9).

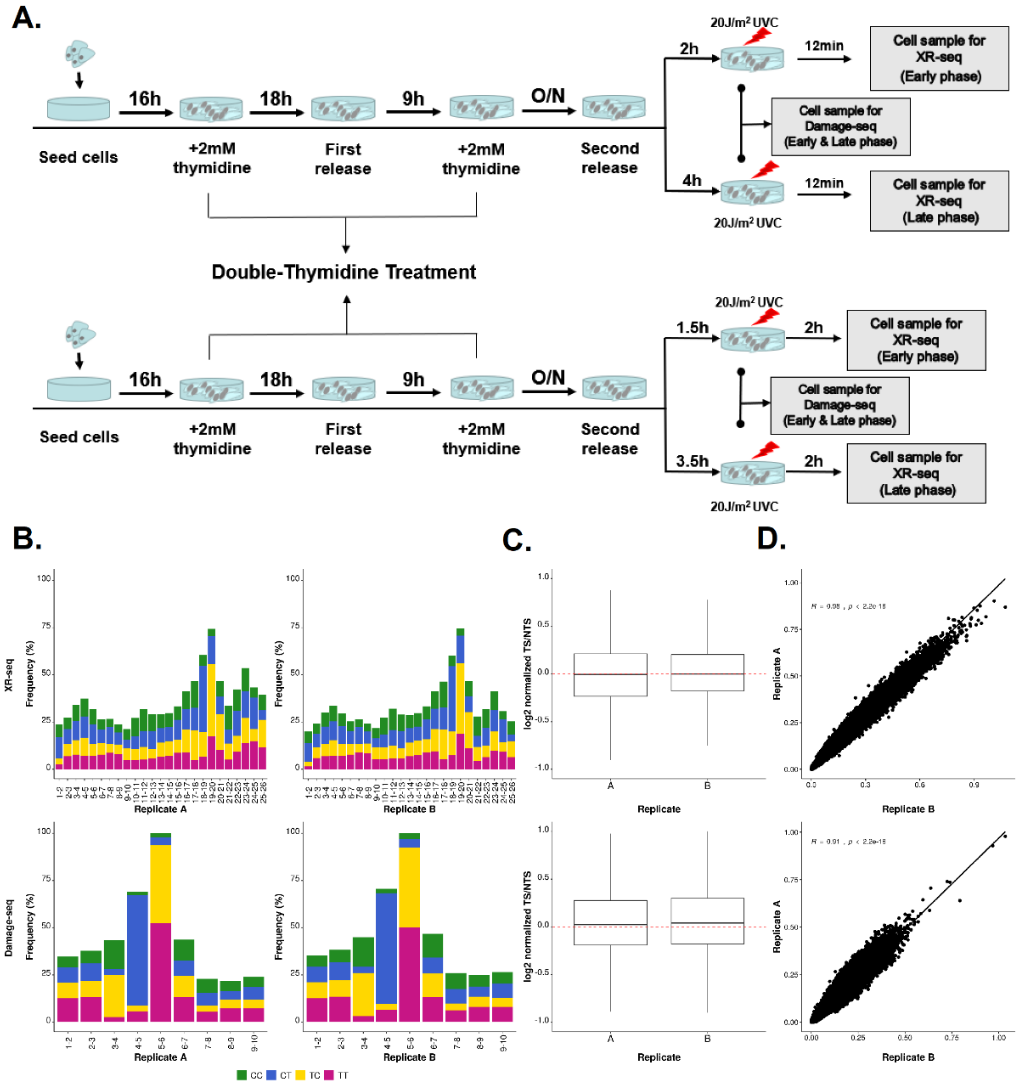

Figure 4.1 A) Experimental setup. B-D) Control figures of (6-4)PP early phased samples at 12 minutes. B) The dinucleotide composition frequency of replicate A and B, respectively. C) log2 transformed TS/NTS ratios of replicate A and B. Row 1 is the results of XR-seq samples, and row 2 is the results of Damage-seq samples. D) The correlation plot of the biological replicates (A & B). Correlation coefficient is calculated by Spearman's rank correlation test.

4.2 Early replication domains are repaired more efficiently than late

replication domains, however, the repair rate of late replication domains elevates while replication proceeds.

To determine how excision repair rates are influenced by replication domains during replication, we compared repair efficiency of early replication domains (ERDs) and late replication domains (LRDs). We obtained replication domains of HeLa cells from a study where a supervised method called Deep Neural Network-Hidden Markov Model was developed to define replication domains from Repli-seq data (Liu et al., 2016). We mapped damage and repair events to corresponding replication domains. To eliminate the effect of a potential bias in damage formation, we normalized repair quantities (XR-seq) by the captured damage events (Damage-seq) in each genomic window (Figure 2A). This approach enabled us to assess the efficiency of repair per damage at a given region, which we refer to as repair rate. Based on an analysis with a Hi-C dataset, the human genome was classified into A/B compartments, which are associated with open and closed chromatin regions, respectively (Lieberman-Aiden et al., 2009). Recently, it was also shown that ERDs and LRDs strongly correlate with A/B compartments respectively (Pope et al., 2014; Ryba et al., 2010). Because ERDs are correlated to open chromatins, these regions are more reachable for excision repair machinery than LRDs. Expectedly, repair rates of ERDs are elevated at the center and gradually reduced towards flanking sites, while LRDs exhibit an opposite pattern (Figure 2A, S18-23). These results suggest that ERDs and their flanking regions are efficiently repaired, whereas less reachable LRDs are poorly repaired. Moreover, LRDs are known to contain higher mutation frequency than other regions (Lawrence et al., 2013; Stamatoyannopoulos et al., 2009), hence; low repair rate of UV damages located at LRDs might be a key factor of mutagenesis in melanoma cancers. On the other hand, the difference between early and late S phases indicates that repair rate is elevated in favor of LRDs when replication timing moves from early to late (Figure 2A-B, S24-29). This time dependent increase in the repair rate of LRDs can be caused by the unfolding of heterochromatin during replication. With the unfolding of the chromatin, more LRD regions will be accessible where the DNA lesions can be recognized and removed by nucleotide excision repair. Also we observe a reduction of repair rate at ERDs, however this reduction might be caused by the relativity of the XR-seq method; increased repair rate at LRDs results in a relative decrease in the repair rate at ERDs, even if repair rate does not quantitatively change at ERDs. In addition, (6-4)PP repair at 12 minutes exhibits minor differences between early and late S phases (Figure 2), potentially because of its fast repair after the damage occurrence (Hu, Adebali, Adar, & Sancar, 2017). Conversely, CPD repair rate at 12 minutes and 2 hours demonstrate significant increase for LRDs and decrease for

ERDs (Figure 2B, p-values < 2.2e-16). Comparison of 12 minutes to 2 hours shows that the repair rate bias is less prominent for later repair. It takes longer time to repair CPDs relative to (6-4)PPs.

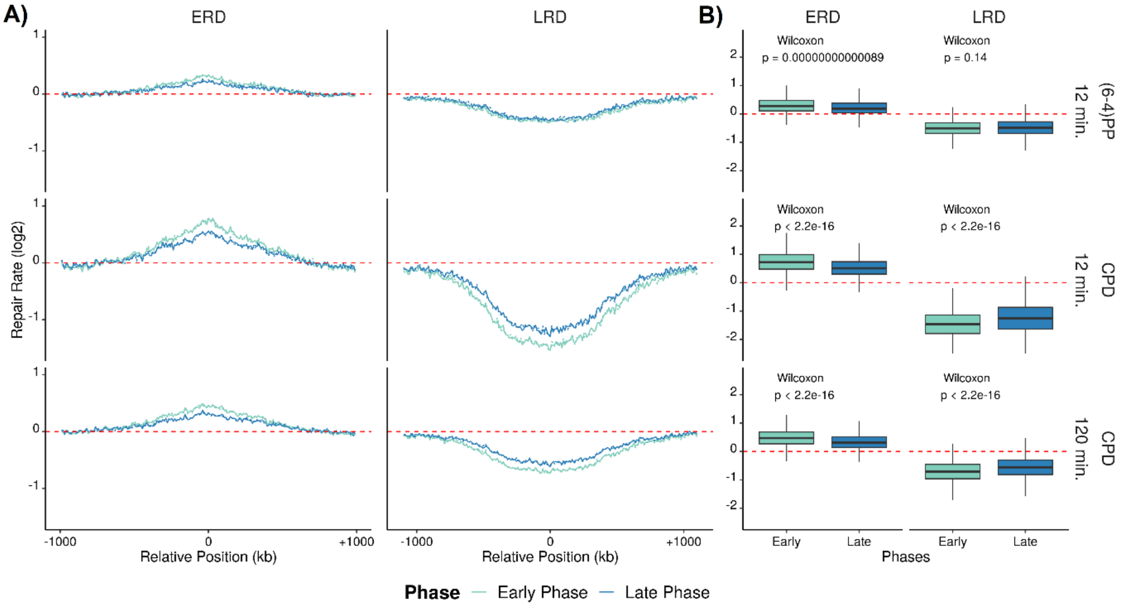

Figure 4.2 The shift of repair efficiency at replication domains during replication timing. A) Repair rates (XR-seq/Damage-seq) are calculated and log2 transformed in 2 Mbp regions with 10 kb intervals, which early replication domains (ERDs, left) and late replication domains (LRDs, right) positioned at the center of the region. B) RPKM values of XR-seq samples are divided by Damage-seq samples (Repair Rate) for both ERDs (left) and LRDs (right) and log2 transformed. Wilcoxon test is used to assess the significance of difference between early and late S phases. The light blue lines are the early phase repair rate values and dark blue lines are the late phase repair rate values. Above the red horizontal dashed line demonstrates that repair is higher than damage, below demonstrates that damage is higher. Analysis is performed on replicate A.

4.3 Variety of chromatin states are associated with dominant repair.

Active chromatin states are repaired effectively, basically because those regions are more accessible to nucleotide excision repair (Adar, Hu, Lieb, & Sancar, 2016). We addressed how repair machinery in ERDs, and LRDs is differentially influenced by the chromatin states during the replication. We retrieved chromatin states of HeLa cells segmented by ChromHMM from UCSC website (Ernst & Kellis, 2017). We intersected the chromatin states with replication domains and mapped damage and repair reads to those regions, for each chromosome. After calculating the repair rates (Figure 3A, S13A-17A), we further assessed early S phase repair relative to late S phase (early/late repair/damage) to observe the replication timing differences in efficiency in the function of chromatin states (Figure 3B, S13B-17B). Generally, repair efficiency is higher in the active chromatin states such as promoters and strong enhancers, which is in agreement with the previous studies (Adar et al., 2016; Hu, Lieb, Sancar, & Adar, 2016). Those regions sustain high repair rates, even at LRDs during the early S phase, that should be condensed and harder to reach (Figure 3A). On the other hand, all the transcription associated chromatin states together with “FaireW” and “Low” chromatin states are highly affected by the replication timing (Figure 3B). “FaireW” represents the regions that are associated to the regulatory activities (Giresi, Kim, McDaniell, Iyer, & Lieb, 2007), whereas “Low” stands for low activity regions that neighboring active sites. On ERDs, although both chromatin states have relatively low repair at early and late S phases, they demonstrate a drastic increase when replication proceeds from early to late. However, on LRD regions, it is hard to make any assumptions considering the wide interquartile range of boxplots of many chromatin states (Figure 3B).

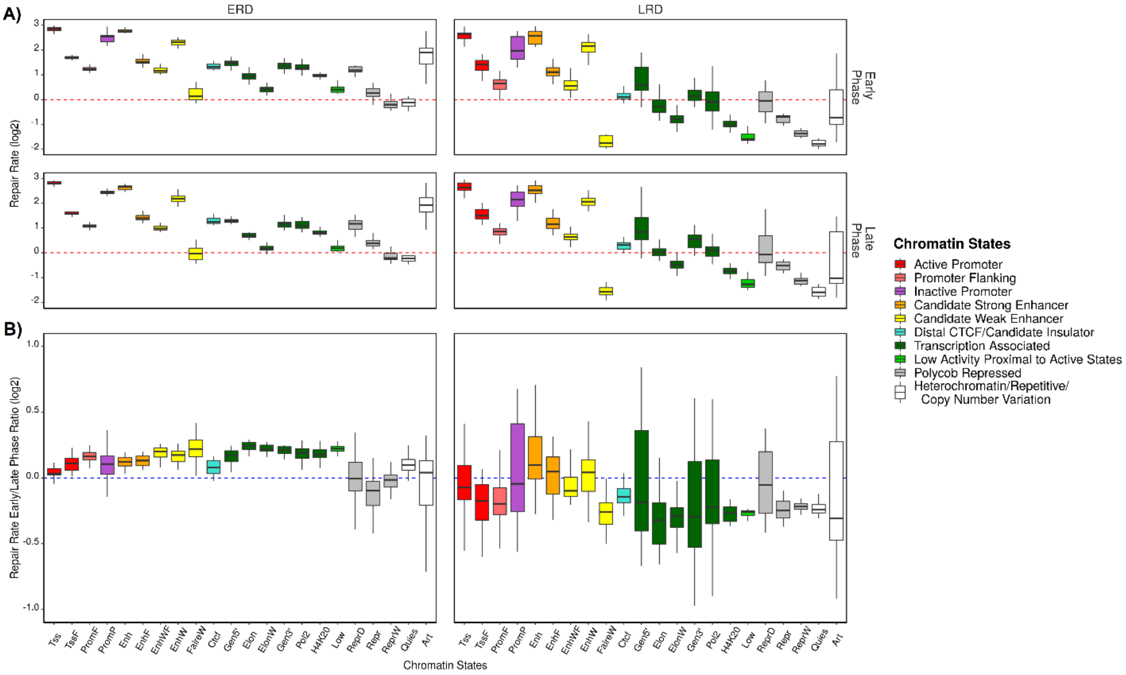

Figure 4.3 The effect of Chromatin States to repair efficiency of replication domains. A) Repair rates (XR-seq/Damage-seq) of CPD samples at 12 minutes are calculated, log2 transformed, B) and for every region, the repair rates at early S phase divided by repair rates at late S phase to spot the chromatin states that are repaired dominant at a phase. Above the red horizontal dashed line demonstrates that repair is higher than damage (A), and the blue horizontal dashed line demonstrates that the chromatin state has higher repair efficiency at early S phase than it has at late S phase (B). Analysis is performed on replicate A.

4.4 ORIs display distinct melanoma mutation counts and strand asymmetry based on their replication domains.

Replication domains are 1 to 2 Mb-sized DNA chunks that involves many small ORIs. Even though analyzing replication domains can exhibit the association of nucleotide excision repair and replication timing on a greater scale, the association of these small ORIs and nucleotide excision repair cannot be explained by only using replication domains. Therefore, we retrieved two independent datasets that are derived from two different methods: okazaki fragment sequencing (OK-seq) and short nascent strand sequencing (SNS-seq). OK-seq quantifies the replication initiation zones that are the sets of closely positioned ORIs using highly purified Okazaki fragments (Petryk et al., 2016), whereas SNS-seq can precisely identifies individual ORIs (Besnard et al., 2012; Langley, Graf, Smith, & Krude, 2016). Using these datasets together with a melanoma mutation dataset that we retrieved from the International Cancer Genome Consortium (ICGC) data portal (see Methods), we examined the mutation profiles at the sites of ORIs where the replication initiates. Because nucleotide excision repair is highly associated to melanoma cancers, we argued that this relation must be reflected to the mutation counts of melanoma. For both OK-seq and SNS-seq data, we assorted the regions based on their corresponding replication domains to detect how mutation profiles of ORIs affected by the domains they are located. Then, we counted the mutations on these regions that are centered at individual ORIs (SNS-seq data, Figure 4A) or initiation zones (OK-seq data, Figure 4B-C) and normalized the mutations with cytosine counts in each bin, to eliminate any bias in potential mutation sites. Also, we gradually extended the region length of initiation zones from 20 kb to 200 kb (Figure 4B-C) for observing the replication effect on a range of scales. Mutation counts of initiation zones differ depending on the replication domains they are located. In agreement with previous studies (Lawrence et al., 2013; Schuster-Bockler & Lehner, 2012; Stamatoyannopoulos et al., 2009), the mutation counts of initiation zones at LRDs elevate, while ERDs contain the initiation zones with the lowest mutation counts (Figure 4B-C). These differences that are related to the replication domains are also persistent for the individual ORIs (Figure 4A). Furthermore, initiation zones at up (UTZs) and down transition zones (DTZs), which are the domains that connect ERDs to LRDs, have mutation counts in between of ERDs and LRDs. Moreover, the flanking sites of initiation zones at transition zones that are close to ERDs have lower counts, whereas the sites that are close to LRDs have higher (Figure 4C, left). Mutation counts not only exhibit a replication domain related difference, but also reveal a strand asymmetry around the initiation zones (Figure 4B-C). The asymmetry suggests that lagging strand (minus strand at left direction; plus strand at right direction) have more mutations than leading, independent of the replication domains. While the initiation zones at LRDs show an explicit strand asymmetry compared to the initiation zones at ERDs, the initiation zones at ERDs have a wider strand asymmetry than that of LRDs. One possible reason can be the amount of ORIs they harbor; earlier studies suggest that ERDs contain significantly higher number of replication origins (Besnard et al., 2012), and the cumulative effect of these ORIs can create a strand asymmetry that is visible on a wider region. Additionally, replication fork movement at LRDs (1.5–2.3 kb/min) is faster than it is at ERDs (1.1–1.2 kb/min) (S. Takebayashi et al., 2005), which can cause more mutation and increased asymmetry between strands. Conversely, individual ORIs obtained from SNS-seq data do not show an explicit strand asymmetry (Figure 4A).

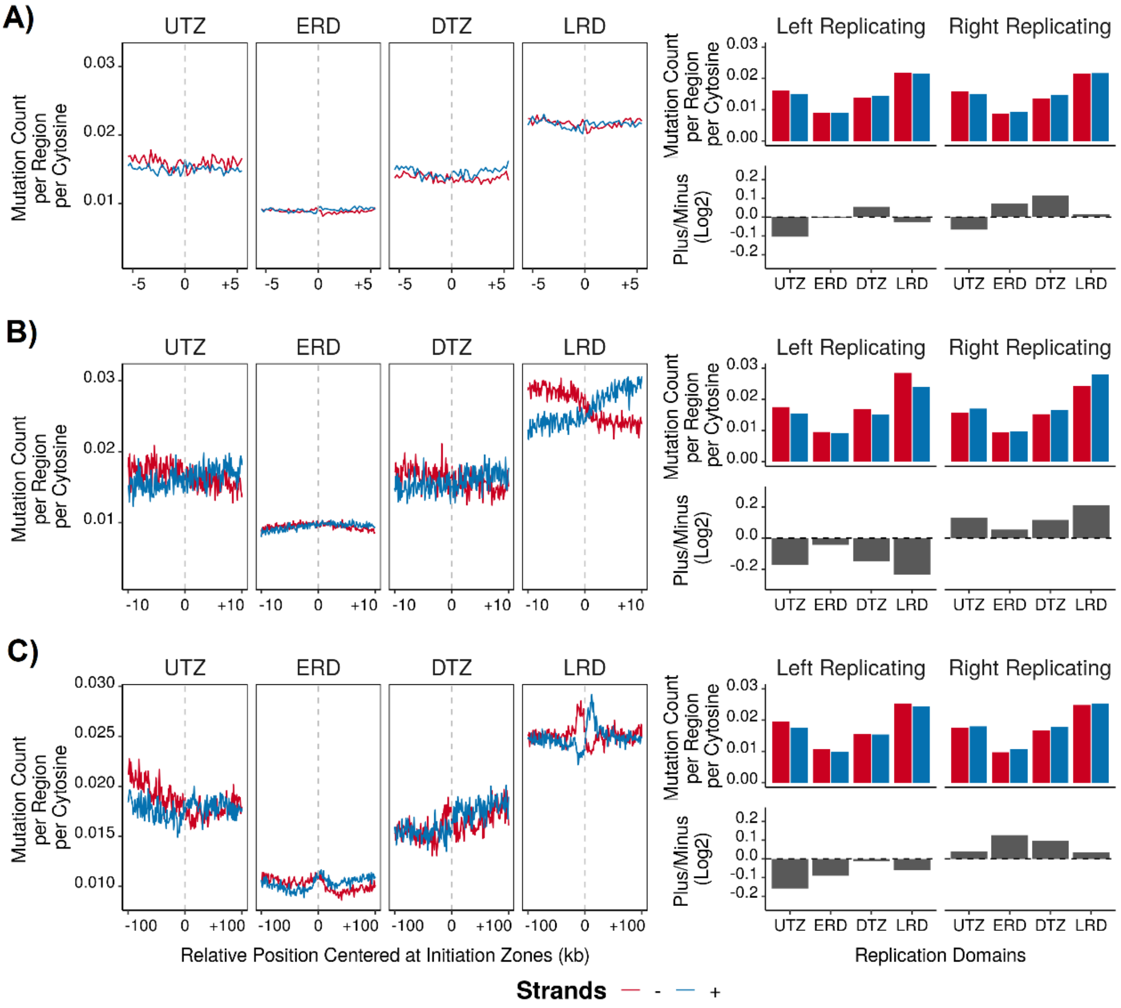

Figure 4.4 Tumor mutation profiles around replication origins and initiation zones for each replication domain. A) C to T mutations are mapped to Replication Origins (SNS-seq) and counted in 10 kb regions with 100 base pair intervals. C to T mutations are mapped to Initiation Zones (OK-seq) B) counted in 20 kb regions with 100 base pair intervals, C) and counted in 200 kb regions with 1000 base pair intervals. Counts are normalized by the number of regions and cytosine counts of each region. Red lines are the plus strands and blue lines are the minus strands. Gray vertical dashed line shows the center of the region. Upper right part demonstrates the strand differences at left (left part of the gray line) and right (right part of the gray line) replicating directions by taking the mean of the intervals, separately for the strands. Below that, strands are divided to each other (Plus/Minus) and log2 transformed to better visualize the asymmetry at each replication domain.

4.5 Asymmetric damage around initiation zones causes asymmetric

repair profiles.

To find whether there is a strand asymmetry at repair and damage profiles similar to melanoma mutations, we mapped repair and damage events to initiation zones independently. Interestingly, A strand asymmetry around the initiation zones is visible for both repair and damage profiles (Figure 5). The asymmetry suggests that lagging template strand harbors more damages and attracts more repair accordingly. Reasoning that nucleotide composition of initiation zones might contribute to the strand asymmetry, we decided to simulate damage and repair signals. We simulated signals via Art simulator (W. Huang, Li, Myers, & Marth, 2012), filtered the signals that resembled the real signals in nucleotide composition, and mapped the filtered ones to the human genome. The simulated signals indeed display a similar strand asymmetry, indicating the contribution of nucleotide composition of the genomic regions surrounding the initiation zones. Nonetheless, simulated signals have lower RPKM values in general. Although nucleotide composition of these signals and real ones are similar, the real repair and damage events occurred at other regions. Therefore, it is expected to observe lower RPKM values for simulated signals.

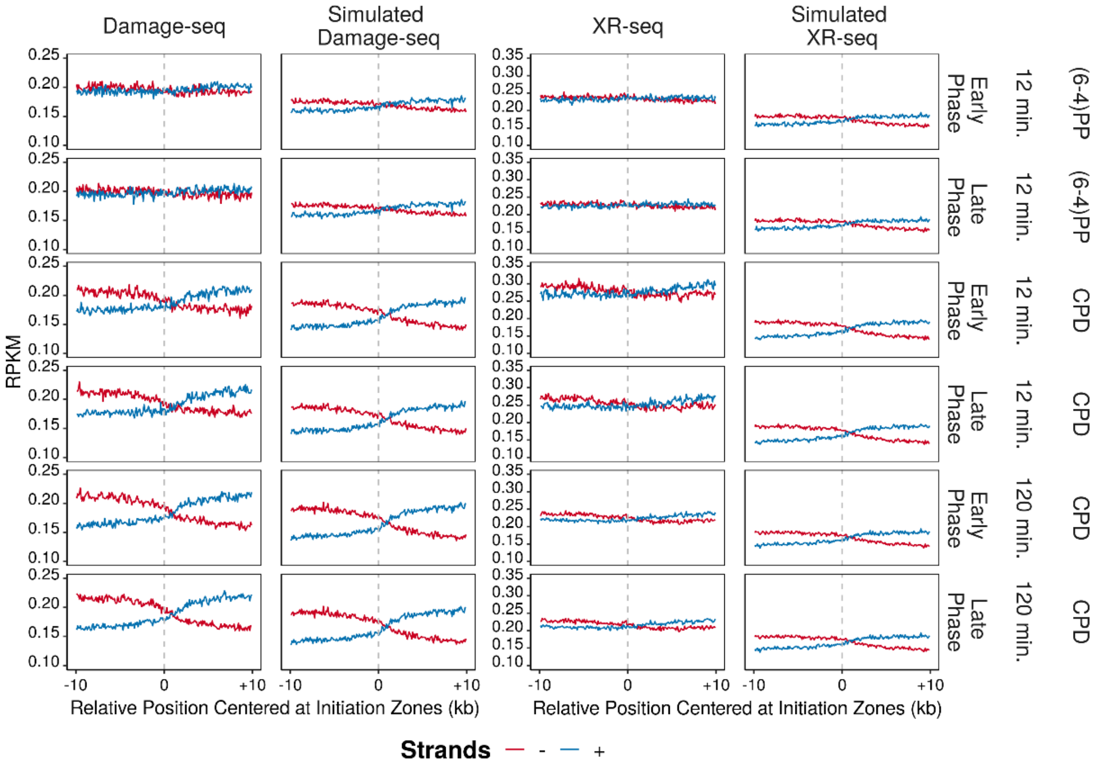

Figure 4.5 Strand asymmetry around initiation zones caused by sequence content. RPKM values of real and simulated Damage-seq samples (left) and XR-seq samples (right) are calculated in 20 kb windows with 100 base pair intervals, which Initiation Zones are positioned at the center of the region. Blue lines represent the plus strands and red lines represent the minus strands. Gray vertical dashed line shows the center of the region.

4.6 Strand asymmetry of excision repair rate

After observing strand asymmetry at mutation counts of melanoma, and damage and repair events independently, we examined the repair rates of (6-4)PP and CPD samples around initiation zones for an asymmetry. Interestingly, a strand asymmetry that is inversely correlated with mutation counts of melanoma is prominent among the CPD damages (Figure 6). The asymmetry indicates an efficient repair of leading strand, which is in agreement with the mutation counts that displayed low mutation on leading strand. This pattern becomes more explicit at a wider scale (Figure S36,37). In addition, CPD samples at 2 hours have elevated asymmetry at a wider scale compared to the CPD samples at 12 minutes (Figure 6, S36, 37, 42-45), because CPDs can be effectively repaired 1 hour after the damage formation. On the contrary, samples with (6-4)PP damages are not showing any strand difference, because of the fast repair ability of nucleotide excision repair for these photo-products. Although we do not observe a distinct mutational strand asymmetry at the individual ORIs, we analyzed the damage and repair events individually, and repair rates around ORIs. Surprisingly, we observed a strand asymmetry at individual ORIs (Figure S46-53, 58-61). While the samples at 2 hours demonstrates an efficient repair at leading template, samples at 12 minutes display an asymmetry that favors lagging template repair. Even though replication forks often move bidirectionally, at some regions, forks tend to move continuously in one direction, either by the moving long distances as a single fork, or multiple forks that are fired simultaneously (S. I. Takebayashi et al., 2017). To observe the effect of replication fork movement, we decided to use the regions that replication forks move in one direction. We retrieved a data which is produced by OK-seq and contains regions that are dominantly replicated in a single direction, termed high replication fork directionality (RFDs) (Petryk et al., 2016). Repair rates at high RFDs display a decrease at the direction of replication fork on wider regions (Figure S62-67). This decrease can be caused by the replication fork itself, considering that the regions replication fork had passed opens and becomes reachable to nucleotide excision repair, while the downstream will be relatively condensed. Also, CPDs at 2 hours demonstrate a visible strand asymmetry at both directions in favor of leading template strand (Figure S62-67). These results suggest that the unidirectional movement of replication fork creates a strand difference by synthesizing one strand as lagging and the opposite as leading for long regions because leading strand is repaired more efficiently.

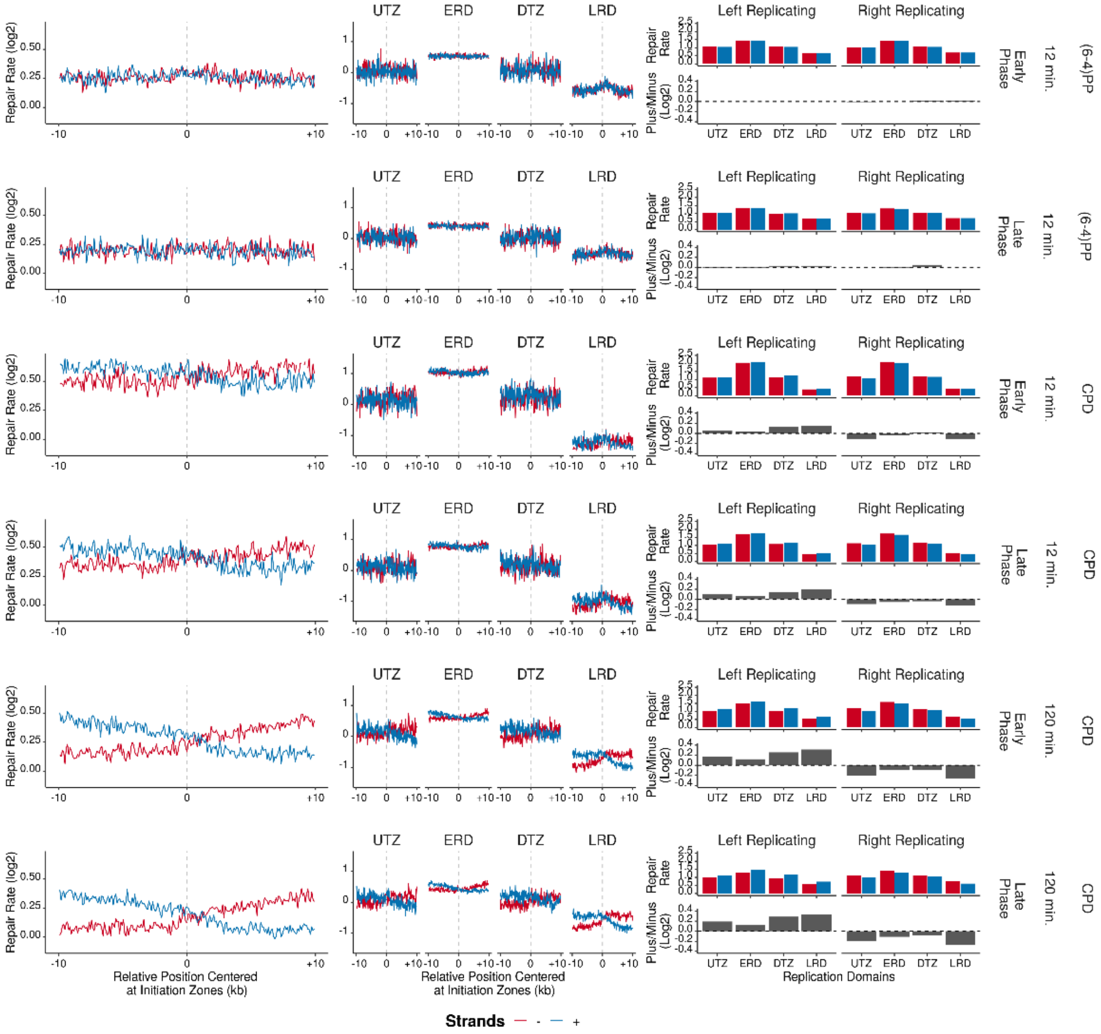

Figure 4.6 Repair rate asymmetry around initiation zones and replication domains. (Left) Repair rates (XR-seq/Damage-seq) are calculated and log2 transformed in 20 kb windows with 100 base pair intervals, which Initiation Zones are positioned at the center of the region. (Middle) Same analysis performed, however initiation zones separated into their corresponding replication domains. (Right) The strand differences at left (left part of the gray line) and right (right part of the gray line) replicating directions are shown by taking the mean of the intervals, separately for the strands. Below that, strands are divided to each other (Plus/Minus) and log2 transformed to better visualize the asymmetry at each replication domain. Blue lines are the plus strands and red lines are the minus strands. Gray vertical dashed line shows the center of the region.

5. DISCUSSION

DNA replication is a highly conserved and regulated temporal process that is essential to genome inheritance. Yet, stochastic effects of DNA replication might cause mutagenesis contributing to cancer (Tomasetti & Vogelstein, 2015). Therefore, an accurate and properly coordinated DNA replication is needed to prevent errors and to preserve DNA fidelity which is constantly threatened by both endogenous and exogenous sources during DNA replication. Considering 70,000 lesions occur in a single cell per day (Lindahl & Barnes, 2000), DNA lesions must be removed before the next cell division, to avoid their permanent conversion into mutations.

On the other hand, DNA excision repair mechanisms are known to relentlessly coup with DNA damages that are potential sites of mutations. In deficiencies of mismatch repair and nucleotide excision repair, there are specific mutational signatures associated which contribute to different cancer types (Helleday, Eshtad, & Nik-Zainal, 2014). Nucleotide excision repair associated signature 7 displays replication timing differences and replication related strand asymmetry (Tomkova et al., 2018). Furthermore, because early replication domains are more reachable relative to late replication domains, mismatch repair causes a mutation difference between these domains by effectively repairing the mismatches at early replication domains (Supek & Lehner, 2015). Similarly, TCR creates a transcriptional strand asymmetry by repairing adducts only at transcribed strands and leaving the opposite strand untouched (Zheng et al., 2014). Even though signature 7 is linked with DNA replication timing and strand asymmetry (Tomkova et al., 2018), the contribution of nucleotide excision repair to mutation differences during replication is still unclear. In this study, we performed Damage-seq and XR-seq methods on UV-irradiated HeLa cells that are synchronized at early and late S phases to quantify (6-4)PP and CPD damages and their repair events.

5.1 DNA replication elevates local nucleotide excision repair by

mediating chromatin opening

In the first part of the study, we examined the repair rate of nucleotide excision repair at large regions, while replication moves from early S phase to late S phase. To examine the repair rate on replication domains, we mapped damage and repair events to these regions. After calculating the repair rates, we found that ERD is repaired faster more efficiently that LRD. Because ERD usually corresponds to open chromatin sites, it is more reachable for nucleotide excision repair, in turn, promotes efficient repair. This result suggests that, like mismatch repair (Supek & Lehner, 2015), nucleotide excision repair creates a mutational difference between ERD and LRD by efficiently repairing ERD. This phenomenon is less detectable for (6-4)PP damages than that of CPDs due to fast repair of (6-4)PPs.

Then, we examined the differences in repair rates of chromatin states for both ERD and LRD. The chromatin states that are associated with close regions are more varying in ERD, whereas the chromatin states that are associated with open regions displays same phenomenon in LRD, simply because ERD have less close regions, while LRD have less open regions. Expectedly, the active promoters and enhancers are repaired efficiently for both ERD and LRD. Moreover, the major difference between phases occurred on close chromatin states. This result indicates that movement of replication elevates the repair rates of close chromatin states by mediating chromatin opening. Lastly, we proposed a simple model to deminstrate the replication effect on repair (FIGUR).

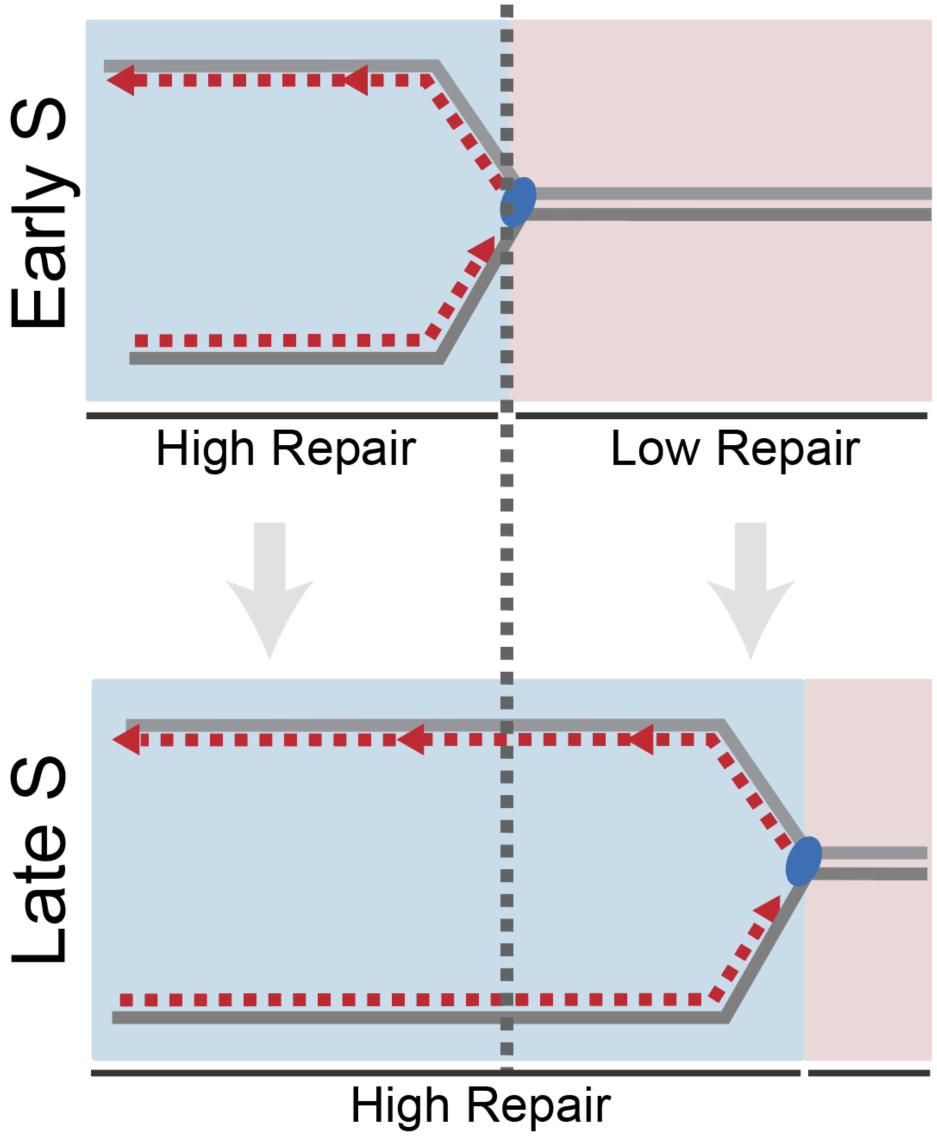

Figure 5.1 Repair preferences of Nucleotide Excision Repair during replication. During the early S phase, open chromatin regions (blue) which mostly correspond to ERDs, are repaired better than the condensed regions (red) because excision repair can reach the open chromatin sites more efficiently than it reaches the condensed regions. While replication continues, those condensed regions loosen and become more reachable for excision repair.

5.2 UV-induced DNA damage, repair and mutagenesis display

replicational strand asymmetry

Secondly, we examined a possible replicational strand asymmetry of mutagenesis, UV-induced DNA damage, and repair, respectively. After quantifying the melanoma mutations on initiation zones, we observed a significant mutational strand asymmetry that contains more mutations at lagging strands. Reasoning that this asymmetry around initiation zones should be related to nucleotide excision repair, we mapped damage and repair events on these regions separately. Remarkably, the same asymmetry was detectable for both damage and repair. Then, we simulated Damage-seq and XR-seq reads to see if the asymmetry is biased by the sequence context and the asymmetry was also detectable for the simulated reads. The results indicates that indeed the strand asymmetry around initiation zones are caused by the sequence context, initially at the level of UV-induced damages. Because there are more damages at lagging strands, repair also increases at lagging strand to cope with the damages. Conversely, when we normalized the repair with damage events, repair rates indicates an opposite asymmetry, higher repair rate at leading strands. These results suggest that even though repair events are higher at lagging strand due to high damages, the efficiency of repair elevates at leading strands, which also explains the less mutation counts at leading strands.

BIBLIOGRAPHY

Boyce, R. P. & Howard-Flanders, P. (1964). Release of ultraviolet light-induced thymine dimers from dna in e. coli k-12. *Proceedings of the National Academy of Sciences of the United States of America*, *51*(2), 293.

Citterio, E., Rademakers, S., van der Horst, G. T., van Gool, A. J., Hoeijmakers, J. H., & Vermeulen, W. (1998). Biochemical and biological characterization of wild-type and atpase-deficient cockayne syndrome b repair protein. *Journal of Biological Chemistry*, *273*(19), 11844–11851.

Cleaver, J. (1968). Defective repair replication of dna in xeroderma pigmentosum. *nature*, *218*(5142), 652–656.

Cleaver, J. E. & Bootsma, D. (1975). Xeroderma pigmentosum: biochemical and genetic characteristics. *Annual review of genetics*, *9*(1), 19–38.

Cockayne, E. A. (1936). Dwarfism with retinal atrophy and deafness. *Archives of disease in childhood*, *11*(61), 1.

Cockayne, E. A. (1946). Dwarfism with retinal atrophy and deafness. *Archives of disease in childhood*, *21*(105), 52–54.

De Boer, J. & Hoeijmakers, J. H. (2000). Nucleotide excision repair and human syndromes. *Carcinogenesis*, *21*(3), 453–460.

Dimitrova, D. S. & Berezney, R. (2002). The spatio-temporal organization of dna replication sites is identical in primary, immortalized and transformed mammalian cells. *Journal of cell science*, *115*(21), 4037–4051.

Douki, T. & Cadet, J. (2001). Individual determination of the yield of the main uv-induced dimeric pyrimidine photoproducts in dna suggests a high mutagenicity of cc photolesions. *Biochemistry*, *40*(8), 2495–2501.

Drapkin, R., Reardon, J. T., Ansari, A., Huang, J.-C., Zawel, L., Ahn, K., Sancar, A., & Reinberg, D. (1994). Dual role of tfiih in dna excision repair and in transcription by rna polymerase ii. *Nature*, *368*(6473), 769–772.

Farkash-Amar, S., Lipson, D., Polten, A., Goren, A., Helmstetter, C., Yakhini, Z., & Simon, I. (2008). Global organization of replication time zones of the mouse genome. *Genome research*, *18*(10), 1562–1570.

Fousteri, M., Vermeulen, W., van Zeeland, A. A., & Mullenders, L. H. (2006). Cockayne syndrome a and b proteins differentially regulate recruitment of chromatin remodeling and repair factors to stalled rna polymerase ii in vivo. *Molecular cell*, *23*(4), 471–482.

Friedberg, E. C., Walker, G. C., Siede, W., & Wood, R. D. (2005). *DNA repair and mutagenesis*. American Society for Microbiology Press.

Giglia-Mari, G., Miquel, C., Theil, A. F., Mari, P.-O., Hoogstraten, D., Ng, J. M., Dinant, C., Hoeijmakers, J. H., & Vermeulen, W. (2006). Dynamic interaction of ttda with tfiih is stabilized by nucleotide excision repair in living cells. *PLoS Biol*, *4*(6), e156.

Hansen, R. S., Thomas, S., Sandstrom, R., Canfield, T. K., Thurman, R. E., Weaver, M., Dorschner, M. O., Gartler, S. M., & Stamatoyannopoulos, J. A. (2010). Sequencing newly replicated dna reveals widespread plasticity in human replication timing. *Proceedings of the National Academy of Sciences*, *107*(1), 139–144.

Haradhvala, N. J., Polak, P., Stojanov, P., Covington, K. R., Shinbrot, E., Hess, J. M., Rheinbay, E., Kim, J., Maruvka, Y. E., Braunstein, L. Z., et al. (2016). Mutational strand asymmetries in cancer genomes reveal mechanisms of dna damage and repair. *Cell*, *164*(3), 538–549.

Hedglin, M. & Benkovic, S. J. (2017). Eukaryotic translesion dna synthesis on the leading and lagging strands: Unique detours around the same obstacle. *Chemical reviews*, *117*(12), 7857–7877.

Hiratani, I., Ryba, T., Itoh, M., Yokochi, T., Schwaiger, M., Chang, C.-W., Lyou, Y., Townes, T. M., Schübeler, D., & Gilbert, D. M. (2008). Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS Biol*, *6*(10), e245.

Hu, J. & Adar, S. (2017). The cartography of uv-induced dna damage formation and dna repair. *Photochemistry and photobiology*, *93*(1), 199–206.

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

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

Jackson, D. A. & Pombo, A. (1998). Replicon clusters are stable units of chromosome structure: evidence that nuclear organization contributes to the efficient activation and propagation of s phase in human cells. *The Journal of cell biology*, *140*(6), 1285–1295.

Khattak, M. & Wang, S. Y. (1972). The photochemical mechanism of pyrimidine cyclobutyl dimerization. *Tetrahedron*, *28*(4), 945–957.

Kiefer, J. (2007). Effects of ultraviolet radiation on dna. In *Chromosomal Alterations* (pp. 39–53). Springer.

Kielbassa, C., Roza, L., & Epe, B. (1997). Wavelength dependence of oxidative dna damage induced by uv and visible light. *Carcinogenesis*, *18*(4), 811–816.

Klungland, A., Höss, M., Gunz, D., Constantinou, A., Clarkson, S. G., Doetsch, P. W., Bolton, P. H., Wood, R. D., & Lindahl, T. (1999). Base excision repair of oxidative dna damage activated by xpg protein. *Molecular cell*, *3*(1), 33–42.

Koren, A., Handsaker, R. E., Kamitaki, N., Karlić, R., Ghosh, S., Polak, P., Eggan, K., & McCarroll, S. A. (2014). Genetic variation in human dna replication timing. *Cell*, *159*(5), 1015–1026.

Langston, L. D., Indiani, C., & O'Donnell, M. (2009). Whither the replisome: emerging perspectives on the dynamic nature of the dna replication machinery. *Cell Cycle*, *8*(17), 2686–2691.

Lawrence, M. S., Stojanov, P., Polak, P., Kryukov, G. V., Cibulskis, K., Sivachenko, A., Carter, S. L., Stewart, C., Mermel, C. H., Roberts, S. A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, *499*(7457), 214–218.

Lehmann, A. R. (2003). Dna repair-deficient diseases, xeroderma pigmentosum, cockayne syndrome and trichothiodystrophy. *Biochimie*, *85*(11), 1101–1111.

Li, S., Wehrenberg, B., Waldman, B. C., & Waldman, A. S. (2018). Mismatch tolerance during homologous recombination in mammalian cells. *DNA repair*, *70*, 25–36.

Li, W. & Sancar, A. (2020). Methodologies for detecting environmentally induced dna damage and repair. *Environmental and Molecular Mutagenesis*, *61*.

Lujan, S. A., Williams, J. S., Pursell, Z. F., Abdulovic-Cui, A. A., Clark, A. B., McElhinny, S. A. N., & Kunkel, T. A. (2012). Mismatch repair balances leading and lagging strand dna replication fidelity. *PLoS Genet*, *8*(10), e1003016.

Marteijn, J. A., Lans, H., Vermeulen, W., & Hoeijmakers, J. H. (2014). Understanding nucleotide excision repair and its roles in cancer and ageing. *Nature reviews Molecular cell biology*, *15*(7), 465–481.

Mizukoshi, T., Kodama, T. S., Fujiwara, Y., Furuno, T., Nakanishi, M., & Iwai, S. (2001). Structural study of dna duplexes containing the (6–4) photoproduct by fluorescence resonance energy transfer. *Nucleic acids research*, *29*(24), 4948–4954.

Modrich, P. (1997). Strand-specific mismatch repair in mammalian cells. *Journal of Biological Chemistry*, *272*(40), 24727–24730.

Mouret, S., Philippe, C., Gracia-Chantegrel, J., Banyasz, A., Karpati, S., Markovitsi, D., & Douki, T. (2010). Uva-induced cyclobutane pyrimidine dimers in dna: a direct photochemical mechanism? *Organic & biomolecular chemistry*, *8*(7), 1706–1711.

Muftuoglu, M., Selzer, R., Tuo, J., Brosh Jr, R. M., & Bohr, V. A. (2002). Phenotypic consequences of mutations in the conserved motifs of the putative helicase domain of the human cockayne syndrome group b gene. *Gene*, *283*(1-2), 27–40.

Nakayasu, H. & Berezney, R. (1989). Mapping replicational sites in the eucaryotic cell nucleus. *Journal of Cell Biology*, *108*(1), 1–11.

Neill, C. A. & Dingwall, M. M. (1950). A syndrome resembling progeria: A review of two cases. *Archives of disease in childhood*, *25*(123), 213.

Nguyen, H. T. & Minton, K. W. (1988). Ultraviolet-induced dimerization of non-adjacent pyrimidines: A potential mechanism for the targeted- 1 frameshift mutation. *Journal of molecular biology*, *200*(4), 681–693.

O'keefe, R. T., Henderson, S. C., & Spector, D. L. (1992). Dynamic organization of dna replication in mammalian cell nuclei: spatially and temporally defined replication of chromosome-specific alpha-satellite dna sequences. *The Journal of cell biology*, *116*(5), 1095–1110.

Park, H., Zhang, K., Ren, Y., Nadji, S., Sinha, N., Taylor, J.-S., & Kang, C. (2002). Crystal structure of a dna decamer containing a cis-syn thymine dimer. *Proceedings of the National Academy of Sciences*, *99*(25), 15965–15970.

Patrick, M. H. (1977). Studies on thymine-derived uv photoproducts in dna—i. formation and biological role of pyrimidine adducts in dna. *Photochemistry and photobiology*, *25*(4), 357–372.

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

Reijns, M. A., Kemp, H., Ding, J., de Procé, S. M., Jackson, A. P., & Taylor, M. S. (2015). Lagging-strand replication shapes the mutational landscape of the genome. *Nature*, *518*(7540), 502–506.

Rupert, C. S., Goodgal, S. H., & Herriott, R. M. (1958). Photoreactivation in vitro of ultraviolet inactivated hemophilus influenzae transforming factor. *The Journal of general physiology*, *41*(3), 451.

Sancar, A. (2016). Mechanisms of dna repair by photolyase and excision nuclease (nobel lecture). *Angewandte Chemie International Edition*, *55*(30), 8502–8527.

Schreier, W. J., Schrader, T. E., Koller, F. O., Gilch, P., Crespo-Hernández, C. E., Swaminathan, V. N., Carell, T., Zinth, W., & Kohler, B. (2007). Thymine dimerization in dna is an ultrafast photoreaction. *Science*, *315*(5812), 625–629.

Schuster-Böckler, B. & Lehner, B. (2012). Chromatin organization is a major influence on regional mutation rates in human cancer cells. *nature*, *488*(7412), 504–507.

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

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

Selby, C. P. & Sancar, A. (1997b). Human transcription-repair coupling factor csb/ercc6 is a dna-stimulated atpase but is not a helicase and does not disrupt the ternary transcription complex of stalled rna polymerase ii. *Journal of Biological Chemistry*, *272*(3), 1885–1890.

Seplyarskiy, V. B., Akkuratov, E. E., Akkuratova, N., Andrianova, M. A., Nikolaev, S. I., Bazykin, G. A., Adameyko, I., & Sunyaev, S. R. (2019). Error-prone bypass of dna lesions during lagging-strand replication is a common source of germline and cancer mutations. *Nature genetics*, *51*(1), 36–41.

Setlow, R. & Carrier, W. (1964). The disappearance of thymine dimers from dna: an error-correcting mechanism. *Proceedings of the National Academy of Sciences of the United States of America*, *51*(2), 226.

Shinbrot, E., Henninger, E. E., Weinhold, N., Covington, K. R., Göksenin, A. Y., Schultz, N., Chao, H., Doddapaneni, H., Muzny, D. M., Gibbs, R. A., et al. (2014). Exonuclease mutations in dna polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication. *Genome research*, *24*(11), 1740–1750.

Stamatoyannopoulos, J. A., Adzhubei, I., Thurman, R. E., Kryukov, G. V., Mirkin, S. M., & Sunyaev, S. R. (2009). Human mutation rate associated with dna replication timing. *Nature genetics*, *41*(4), 393–395.

Sugasawa, K., Ng, J. M., Masutani, C., Iwai, S., van der Spek, P. J., Eker, A. P., Hanaoka, F., Bootsma, D., & Hoeijmakers, J. H. (1998). Xeroderma pigmentosum group c protein complex is the initiator of global genome nucleotide excision repair. *Molecular cell*, *2*(2), 223–232.

Svejstrup, J. Q. (2002). Mechanisms of transcription-coupled dna repair. *Nature Reviews Molecular Cell Biology*, *3*(1), 21–29.

Takebayashi, S.-i., Ogata, M., & Okumura, K. (2017). Anatomy of mammalian replication domains. *Genes*, *8*(4), 110.

Taylor, J.-S. & Brockie, I. R. (1988). Synthesis of a trans-syn thymine dimer building block. solid phase synthesis of cgtat [t, s] tatgc. *Nucleic acids research*, *16*(11), 5123–5136.

Taylor, J. S. & Cohrs, M. P. (1987). Dna, light, and dewar pyrimidinones: the structure and biological significance to tpt3. *Journal of the American Chemical Society*, *109*(9), 2834–2835.

Tomkova, M., Tomek, J., Kriaucionis, S., & Schuster-Böckler, B. (2018). Mutational signature distribution varies with dna replication timing and strand asymmetry. *Genome biology*, *19*(1), 1–12.

Tornaletti, S., Reines, D., & Hanawalt, P. C. (1999). Structural characterization of rna polymerase ii complexes arrested by a cyclobutane pyrimidine dimer in the transcribed strand of template dna. *Journal of Biological Chemistry*, *274*(34), 24124–24130.

Van Hoffen, A., Venema, J., Meschini, R., Van Zeeland, A., & Mullenders, L. (1995). Transcription-coupled repair removes both cyclobutane pyrimidine dimers and 6-4 photoproducts with equal efficiency and in a sequential way from transcribed dna in xeroderma pigmentosum group c fibroblasts. *The EMBO journal*, *14*(2), 360–367.

Wacker, A., Dellweg, H., Träger, L., Kornhauser, A., Lodemann, E., Türck, G., Selzer, R., Chandra, P., & Ishimoto, M. (1964). Organic photochemistry of nucleic acids. *Photochemistry and Photobiology*, *3*(4), 369–394.

Whitmore, S., Potten, C., Chadwick, C., Strickland, P. T., & Morison, W. (2001). Effect of photoreactivating light on uv radiation-induced alterations in human skin. *Photodermatology, photoimmunology & photomedicine*, *17*(5), 213–217.

Yeeles, J. T., Poli, J., Marians, K. J., & Pasero, P. (2013). Rescuing stalled or damaged replication forks. *Cold Spring Harbor perspectives in biology*, *5*(5), a012815.

Yimit, A., Adebali, O., Sancar, A., & Jiang, Y. (2019). Differential damage and repair of dna-adducts induced by anti-cancer drug cisplatin across mouse organs. *Nature communications*, *10*(1), 1–11.

6. APPENDIX

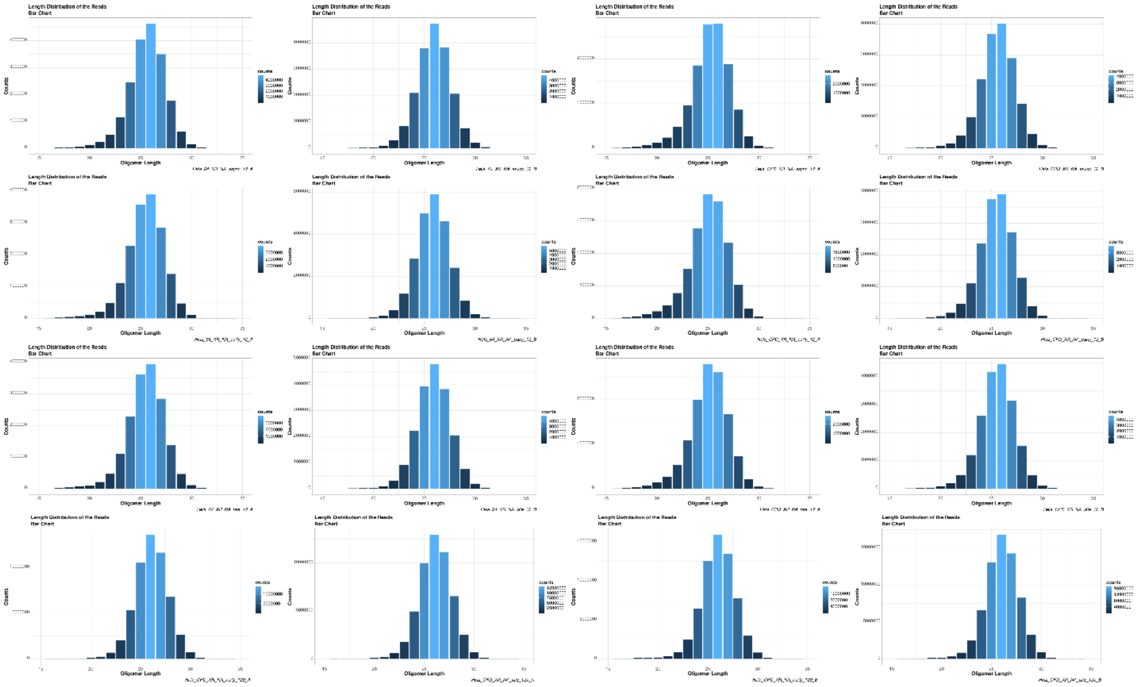

Figure 6.1 Length distribution of excised oligomers of XR-seq samples after adaptor trimming and duplicate removal. Majority of the oligomers are 26 nucleotides long.

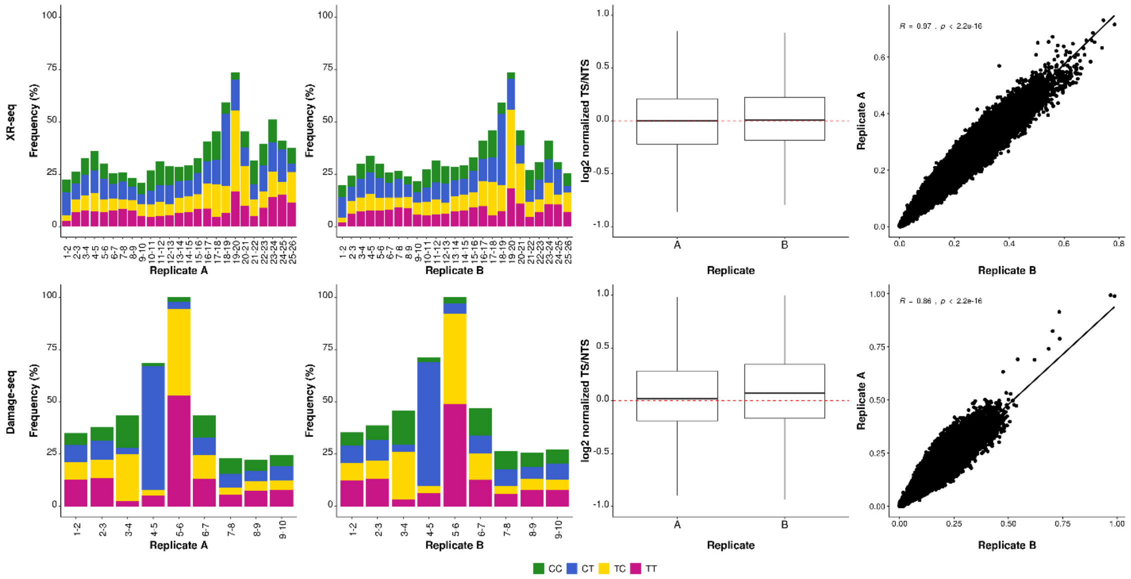

Figure 6.2 Control figures of (6-4)PP asynchronized samples at 12 minutes. Column 1 is the correlation plot of the biological replicates (A & B). Column 1 and 2 displays the dinucleotide composition frequency of replicate A and B, respectively. Column 3 is the log2 transformed TS/NTS ratios of replicate A and B. Row 1 is the results of XR-seq samples, and row 2 is the results of Damage-seq samples. Column 4 is the correlation plot of the biological replicates (A & B). Correlation coefficient is calculated by Spearman's rank correlation test.

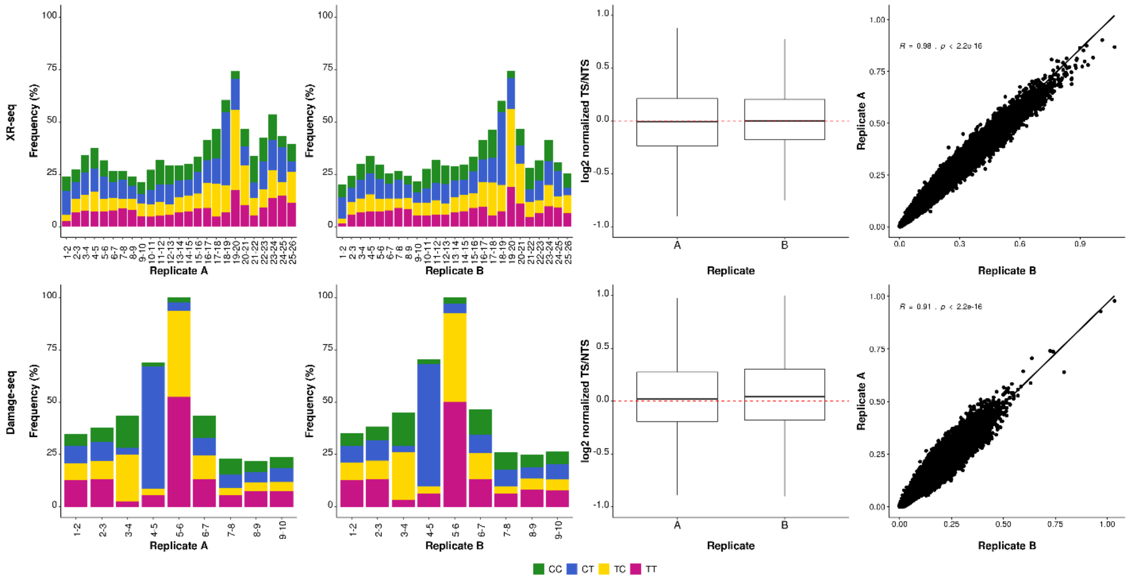

Figure 6.3 Control figures of (6-4)PP early phased samples at 12 minutes. Column 1 is the correlation plot of the biological replicates (A & B). Column 1 and 2 displays the dinucleotide composition frequency of replicate A and B, respectively. Column 3 is the log2 transformed TS/NTS ratios of replicate A and B. Row 1 is the results of XR-seq samples, and row 2 is the results of Damage-seq samples. Column 4 is the correlation plot of the biological replicates (A & B). Correlation coefficient is calculated by Spearman's rank correlation test.

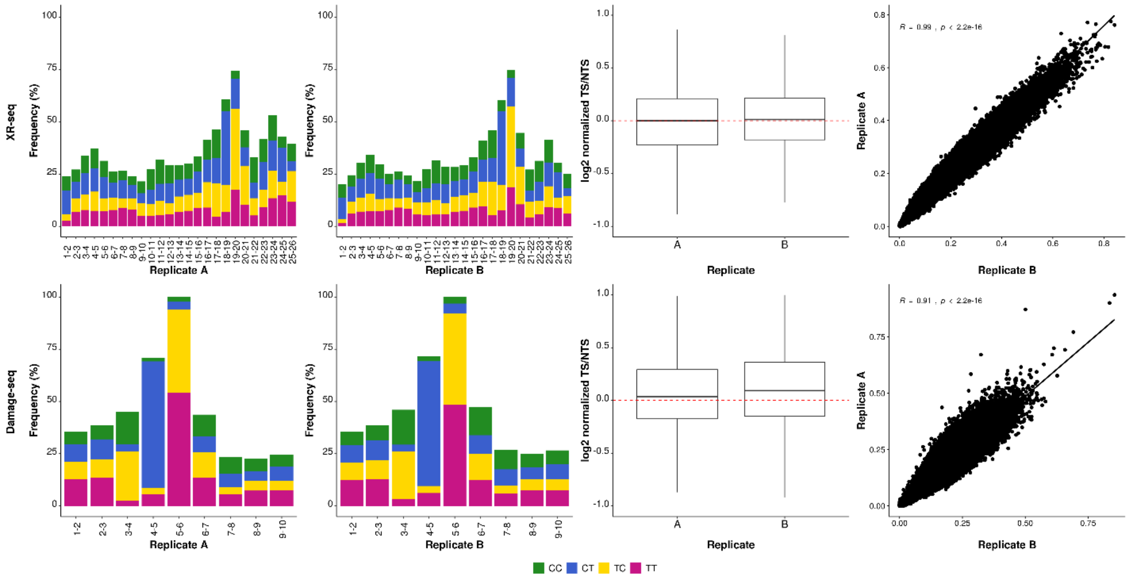

Figure 6.4 Column 1 is the correlation plot of the biological replicates (A & B). Column 1 and 2 displays the dinucleotide composition frequency of replicate A and B, respectively. Column 3 is the log2 transformed TS/NTS ratios of replicate A and B. Row 1 is the results of XR-seq samples, and row 2 is the results of Damage-seq samples. Column 4 is the correlation plot of the biological replicates (A & B). Correlation coefficient is calculated by Spearman's rank correlation test.

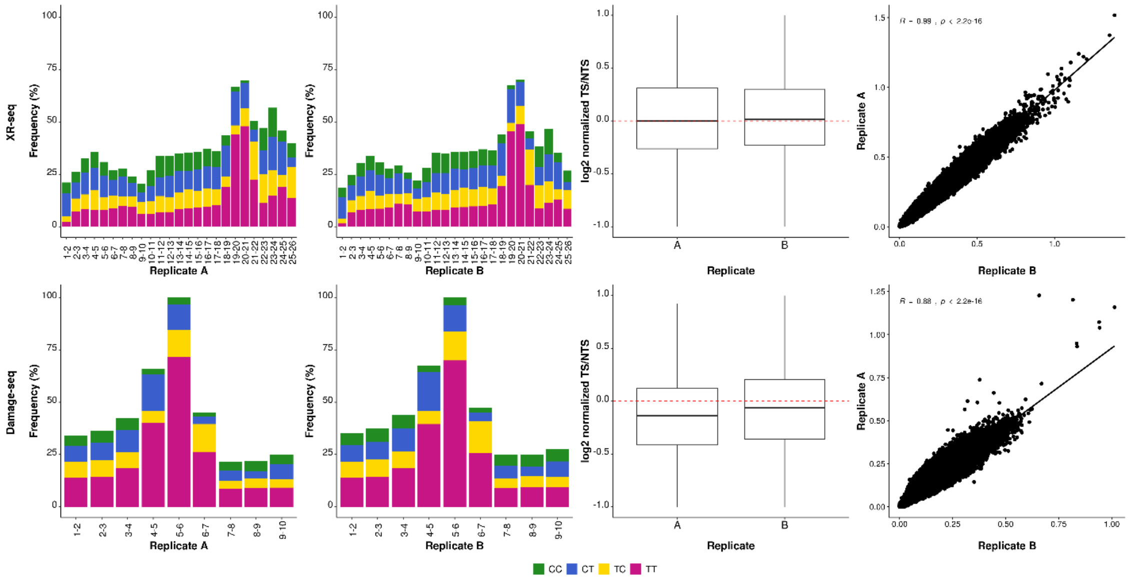

Figure 6.5 Control figures of CPD asynchronized samples at 12 minutes. Column 1 is the correlation plot of the biological replicates (A & B). Column 1 and 2 displays the dinucleotide composition frequency of replicate A and B, respectively. Column 3 is the log2 transformed TS/NTS ratios of replicate A and B. Row 1 is the results of XR-seq samples, and row 2 is the results of Damage-seq samples. Column 4 is the correlation plot of the biological replicates (A & B). Correlation coefficient is calculated by Spearman's rank correlation test.

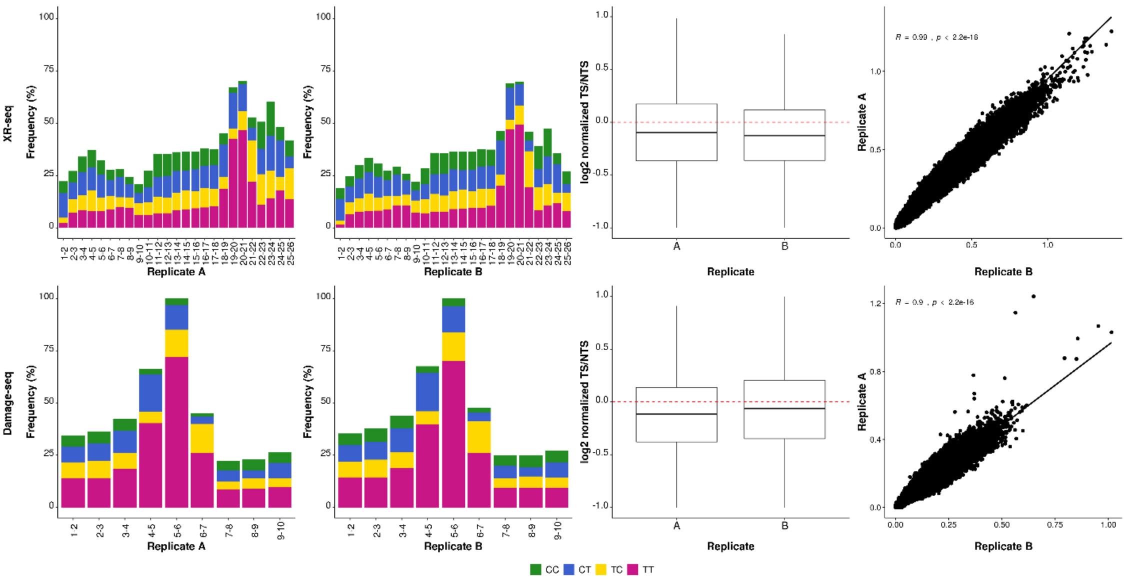

Figure 6.6 Control figures of CPD early phased samples at 12 minutes. Column 1 is the correlation plot of the biological replicates (A & B). Column 1 and 2 displays the dinucleotide composition frequency of replicate A and B, respectively. Column 3 is the log2 transformed TS/NTS ratios of replicate A and B. Row 1 is the results of XR-seq samples, and row 2 is the results of Damage-seq samples. Column 4 is the correlation plot of the biological replicates (A & B). Correlation coefficient is calculated by Spearman's rank correlation test.

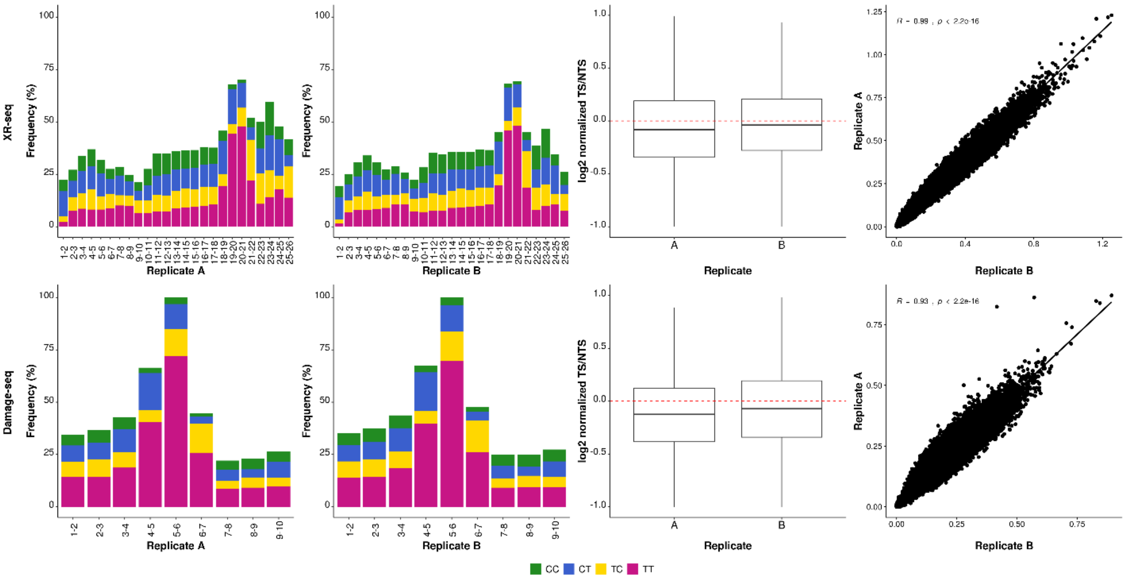

Figure 6.7 Control figures of CPD late phased samples at 12 minutes. Column 1 is the correlation plot of the biological replicates (A & B). Column 1 and 2 displays the dinucleotide composition frequency of replicate A and B, respectively. Column 3 is the log2 transformed TS/NTS ratios of replicate A and B. Row 1 is the results of XR-seq samples, and row 2 is the results of Damage-seq samples. Column 4 is the correlation plot of the biological replicates (A & B). Correlation coefficient is calculated by Spearman's rank correlation test.

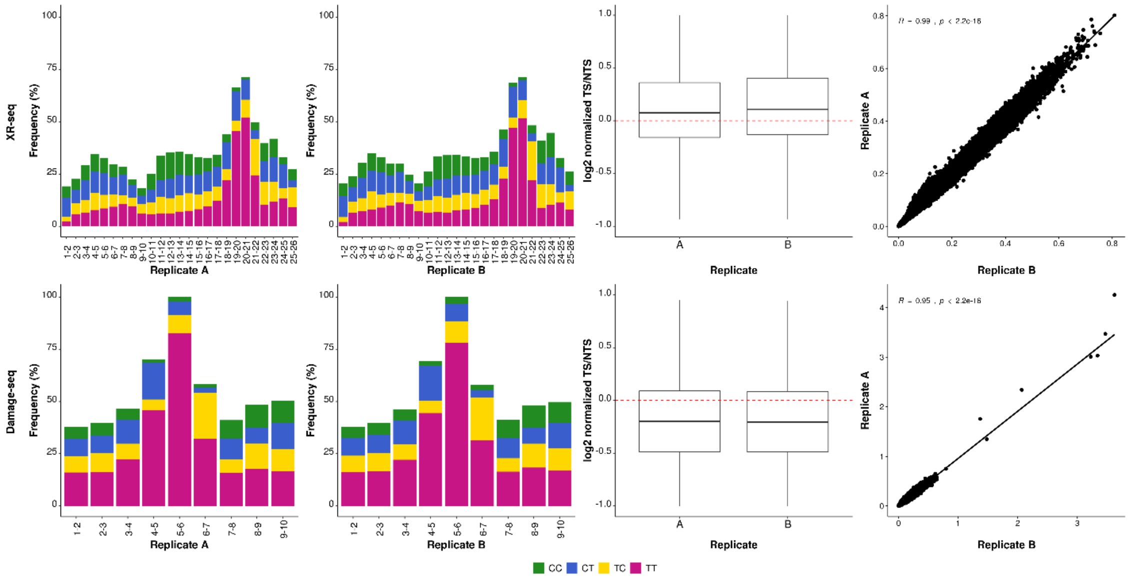

Figure 6.8 Control figures of CPD early phased samples at 120 minutes. Column 1 is the correlation plot of the biological replicates (A & B). Column 1 and 2 displays the dinucleotide composition frequency of replicate A and B, respectively. Column 3 is the log2 transformed TS/NTS ratios of replicate A and B. Row 1 is the results of XR-seq samples, and row 2 is the results of Damage-seq samples. Column 4 is the correlation plot of the biological replicates (A & B). Correlation coefficient is calculated by Spearman's rank correlation test.

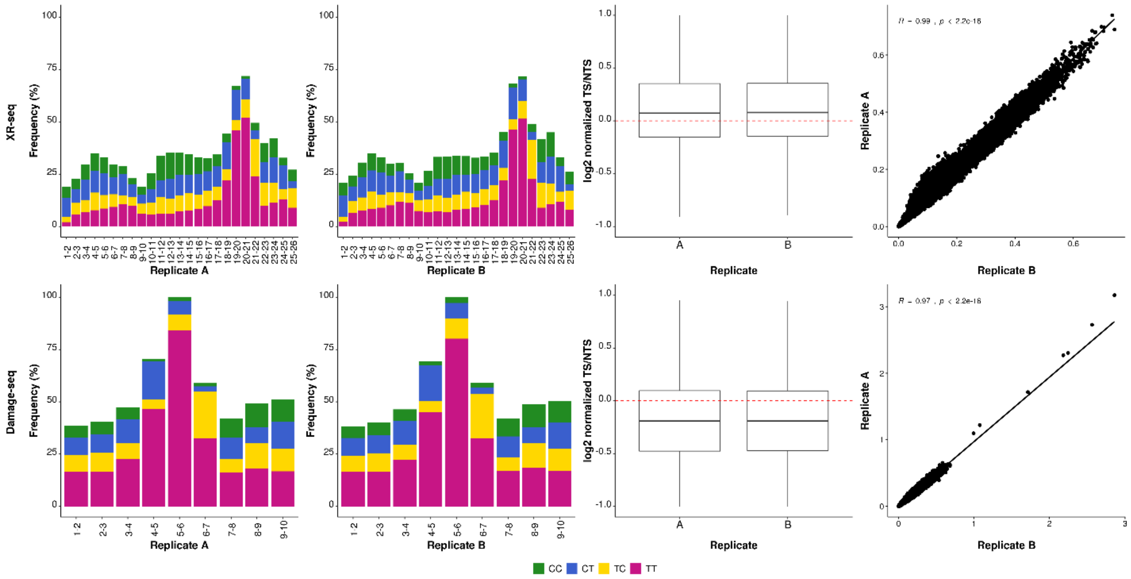

Figure 6.9 Control figures of CPD late phased samples at 120 minutes. Column 1 is the correlation plot of the biological replicates (A & B). Column 1 and 2 displays the dinucleotide composition frequency of replicate A and B, respectively. Column 3 is the log2 transformed TS/NTS ratios of replicate A and B. Row 1 is the results of XR-seq samples, and row 2 is the results of Damage-seq samples. Column 4 is the correlation plot of the biological replicates (A & B). Correlation coefficient is calculated by Spearman's rank correlation test.