THE RELATIONSHIP BETWEEN 3D GENOME ORGANIZATION AND UV-INDUCED DNA DAMAGE AND REPAIR

by
ÜMIT AKKÖSE

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

Sabancı University
June 2023

THE RELATIONSHIP BETWEEN 3D GENOME ORGANIZATION AND UV-INDUCED DNA DAMAGE AND REPAIR

Approved by:

Asst. Prof. Ogün Adebali . . . .
(Thesis Supervisor)

Asst. Prof. Onur Öztaş . . . .

Assoc. Prof. Kamer Kaya . . . .

Date of Approval: June 21, 2023

ABSTRACT

THE RELATIONSHIP BETWEEN 3D GENOME ORGANIZATION AND UV-INDUCED DNA DAMAGE AND REPAIR

ÜMIT AKKÖSE

Molecular Biology, Genetics And Bioengineering M.Sc. THESIS, JUNE 2023

Thesis Supervisor: Asst. Prof. Ogün Adebali

Keywords: UV, DNA damage, DNA repair, 3D genome, NGS simulation

The three-dimensional (3D) configuration of the eukaryotic genome is essential for myriad cellular processes, such as the modulation of gene expression and the orchestration of epigenetic regulation, as well as the preservation of genome integrity. Nonetheless, understanding interaction between UV-induced DNA damage and subsequent repair mechanisms, and their connection with the genome's 3D structure remains underexplored. In the present study, we harness Hi-C, Damage-seq, and XR-seq datasets, complemented by in silico simulations, to look into the interconnection between UV damage and the 3D genome organization. Our findings reveals that the genome's peripheral 3D configuration acts as a defensive barrier, safeguarding the central genomic DNA sectors from UV-induced damage. Furthermore, we found that potential damage sites of pyrimidine-pyrimidone (6-4) photoproducts appear with a higher frequency in the nucleus center, possibly suggesting an evolutionary selection pressure working against the formation of these sites at the genome periphery. We did not find any correlation between the effectiveness of DNA repair and the 3D structure of the genome 12 minutes after UV radiation, showing that UV radiation changes the 3D organization of the genome in a relatively short time frame. Surprisingly, two hours after UV induction, we detected more proficient repair activity in the nucleus center relative to its periphery. Our results provide valuable insights for the understanding of the etiology of cancer and other diseases. The interplay between UV radiation and the 3D organization of the genome may contribute significantly to the emergence of genetic mutations and genomic instability.

ÖZET

3D GENOM ORGANIZASYONU İLE UV-KAYNAKLI DNA HASARI VE ONARIMI ARASINDAKI İLİŞKİ

ÜMIT AKKÖSE

Moleküler Biyoloji, Genetik ve Biyomühendislik YÜKSEK LİSANS TEZİ,
HAZİRAN 2023

Tez Danışmanı: Asst. Prof. Ogün Adebali

Anahtar Kelimeler: UV, DNA hasarı, DNA onarımı, 3D genom, NGS simülasyonu

Ökaryot genomunun üç boyutlu konfigürasyonu, gen ifadesi, epigenetik düzenleme ve genom bütünlüğünün korunması gibi birçok hücresel sürece esastır. Bununla birlikte, UV sebepli DNA hasarı ve ardından gelen onarım mekanizmaları arasındaki karmaşık etkileşimin anlaşılması ve bunların genomun 3D yapısıyla olan bağlantısı hala yeteri kadar araştırılmamıştır. Bu çalışmada, Hi-C, Damage-seq ve XR-seq veri setleri ve in silico simülasyonlar kullanarak, UV hasarı ve 3D genom organizasyonu arasındaki ilişkiyi inceledik. Bulgularımız, genomun çevresel 3D konfigürasyonunun koruyucu rolünü aydınlatıyor ve bu yapının, UV kaynaklı hasardan merkezi genomik DNA bölgelerini koruyan bir savunma bariyeri olarak işlev gördüğünü ortaya koyuyor. İlginçtir ki, pirimidin-pirimidon (6-4) fotoproduktlarının potansiyel hasar yerlerinin, çekirdek merkezinde daha yüksek frekansta ortaya çıktığını bulduk. Bu keşif, bu bölgelerin genom dış bölgesinde oluşumuna karşı çalışan bir evrimsel seçilim basıncını ima edebilir. UV ışınına maruz kaldıktan 12 dakika sonrasında, DNA onarım verimliliği ve genomun 3D yapısı arasında anlamlı bir korelasyon bulamadık. Bu gözlem, genomun 3D organizasyonunun, nispeten kısa bir süre içinde UV radyasyonuna yanıt olarak hızlı bir değişiklik geçirdiğine işaret ediyor. Ancak, UV uygulamasından iki saat sonra, çekirdek merkezindeki onarım aktivitesinin, dış bölgelerine göre daha iyi olduğunu tespit ettik. Sonuçlarımız, kanser ve diğer hastalıkların etiyolojisinin anlaşılması için değerli içgörüler sağlar. UV radyasyonu ve genomun 3D organizasyonu arasındaki etkileşim, genetik mutasyonların ortaya çıkışına ve genomik istikrarsızlığa önemli ölçüde katkıda bulunabilir.

To my supportive family and friends

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

(6-4)PP: Pyrimidine-pyrimidone (6-4)

CPD: Cyclobutane pyrimidine dimer

DSB: Double-strand break

TAD: Topologically associating domain

TCR: Transcription-coupled repair

Hi-C: Chromosome conformation capture

PCR: Polymerase chain reaction

TFIIH: Transcription factor IIH

NGS: Next generation sequencing

CNV: Copy number variation

RPKM: Reads rer kilobase per million mapped reads

1. INTRODUCTION

In lieu of their linear representations, eukaryotic genomes reside within the cellular nucleus as an intricate three-dimensional (3D) structure. This structure, known as chromatin, comprises DNA and an array of proteins that facilitate the packaging, organization, and regulation of the genetic information ensconced within the genome (Woodcock & Dimitrov, 2001). The 3D structuring of the genome is instrumental to accurate gene expression and regulation. It ensures that select regions of DNA are readily accessible to proteins and enzymes that orchestrate transcription and other cellular processes (Pope, Ryba, Dileep, Yue, Wu, Denas, Vera, Wang, Hansen, Canfield, Thurman, Cheng, Gulsoy, Dennis, Snyder, Stamatoyannopoulos, Taylor, Hardison, Kahveci, Ren & Gilbert, 2014; Sanders, Freeman, Xu, Golloshi, Stallard, Hill, San Martin, Balajee & McCord, 2020; Schwarzer, Abdennur, Goloborodko, Pekowska, Fudenberg, Loe-Mie, Fonseca, Huber, Haering, Mirny & Spitz, 2017). Current research has highlighted the impact of the genome's 3D structure on disease onset and progression, emphasizing the importance of understanding genomic organization in health and disease contexts (Chakraborty & Ay, 2019).

The use of genome-wide chromosome conformation capture (Hi-C) methodologies has revealed key aspects of chromosome 3D organization, including compartmentalization, topologically associating domains (TADs), and loops. Lieberman-Aiden et al. (Lieberman-Aiden, van Berkum, Williams, Imakaev, Ragoczy, Telling, Amit, Lajoie, Sabo, Dorschner, Sandstrom, Bernstein, Bender, Groudine, Gnirke, Stamatoyannopoulos, Mirny, Lander & Dekker, 2009) determined that at the megabase scale, the genome is partitioned into two compartments—dubbed as A and B compartments. Interactions between loci are predominantly confined within these compartments. The A compartment correlates with open chromatin, whereas the B compartment is linked with closed chromatin. On a finer scale, at the sub-megabase level, chromosomes are structured into domains that favor intra-domain interactions over inter-domain interactions with adjacent cis-chromatin domains (Dixon, Selvaraj, Yue, Kim, Li, Shen, Hu, Liu & Ren, 2012; Hou, Li, Qin & Corces, 2012; Nora, Lajoie, Schulz, Giorgetti, Okamoto, Servant, Piolot, van Berkum, Meisig, Sedat, Gribnau, Barillot, Bluthgen, Dekker & Heard, 2012). These contact domains, now widely referred to as TADs (Nora et al., 2012), are seen across a multitude of species, suggesting a conserved characteristic of genome organization. TADs represent a functionally privileged scale of chromosome folding, and the restriction of functional contacts within TADs is vital for the correct regulation of genes. Chromatin looping interactions further contribute to long-range gene regulation by connecting genes to distant regulatory elements via the loop extrusion mechanism (Sanyal, Lajoie, Jain & Dekker, 2012).

Damage-seq utilizes the characteristic stalling of DNA polymerase at lesion sites as a primary mechanism to precisely detect damage locations (Hu, Lieb, Sancar & Adar, 2016). In essence, the Damage-seq methodology can be tailored to identify any type of DNA damage that inhibits the normal functioning of the DNA polymerase, provided that a damage-specific antibody is available. To briefly outline the process, post the induction of damage, genomic DNA undergoes a process of sonication, followed by the ligation to initial primers, and subsequent denaturation. DNA lesions are then selectively immunoprecipitated using damage-specific antibodies, and subsequently enriched. This enrichment is followed by the annealing of a biotinylated primer, which is extended by a specific polymerase known as Q5 DNA polymerase. The Q5 polymerase extends the primer until encountering the DNA lesion, without synthesizing the damaged site. An adapter is then ligated to this extended primer to facilitate its amplification via polymerase chain reaction (PCR). Ultimately, the amplified oligomers are subjected to sequencing and can be analyzed (Figure 1.1A).

XR-seq captures the 22-30 nucleotide long excised oligomers that are produced after the dual incision of the lesion site to measure the repair of DNA damages coordinated by the nucleotide excision repair mechanism (Hu et al., 2016). Following incision, the excised oligomers are immunoprecipitated by the transcription factor IIH (TFIIH), and adapters are ligated from both ends. Subsequently, the oligomers undergo a selection process tailored to the specific DNA damage of interest, which is performed by immunoprecipitation using damage-specific antibodies. The lesions within the remaining oligomers are reversed using photolyases, ensuring a successful PCR amplification process. Finally, these oligomers are subjected to sequencing to derive an in-depth understanding of the DNA damage and repair mechanism (Figure 1.1B).

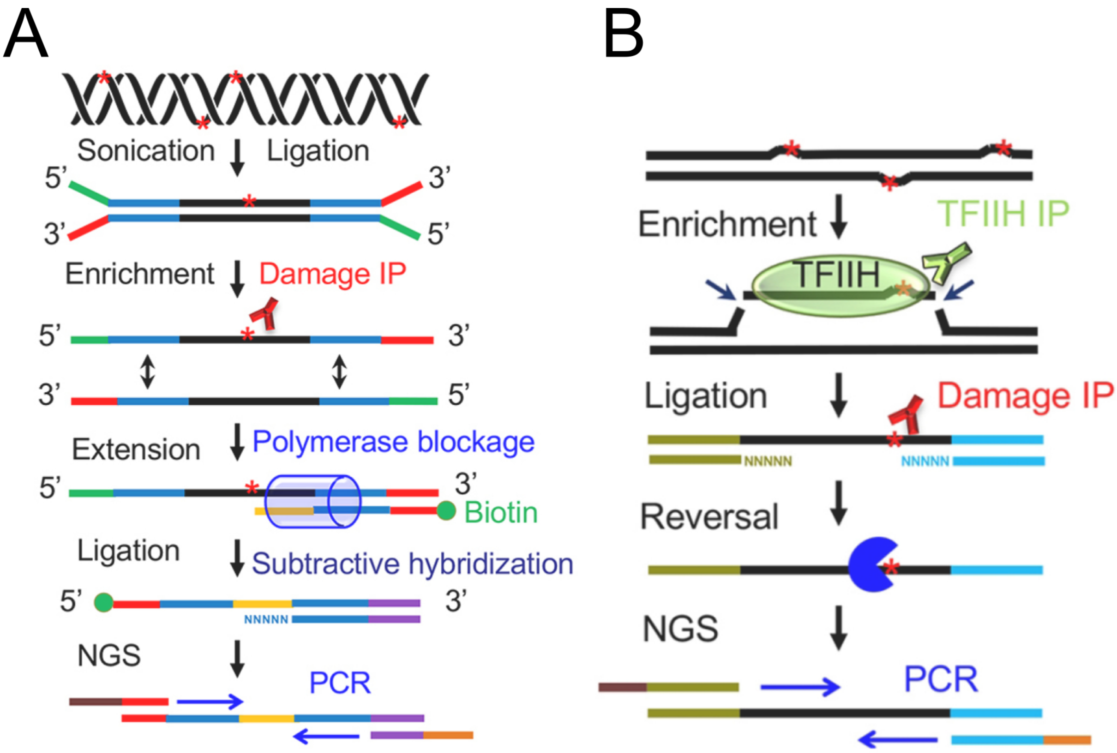

Figure 1.1 Schematic representation of A, Damage-seq and B, XR-seq methods. (Li & Sancar, 2020)

Previous studies have mapped UV and cisplatin-induced DNA damage on human cell lines, uncovering the substantial impact of chromatin states on damage formation (Adar, Hu, Lieb & Sancar, 2016; Hu, Adar, Selby, Lieb & Sancar, 2015; Hu, Adebali, Adar & Sancar, 2017; Mao, Smerdon, Roberts & Wyrick, 2016). The occurrence of UV-induced pyrimidine-pyrimidone (6-4) photoproduct [(6-4)PP] and cyclobutane pyrimidine dimer (CPD) is primarily determined by sequence context and is presumed to be uniform throughout the genome (Hu et al., 2017). Nonetheless, the repair rates for (6-4)PP and CPD are influenced by varying degrees by chromatin states, transcription factor binding, and transcription (Hu et al., 2017). Both damage types are repaired more efficiently in regions of open chromatin and DNaseI hypersensitivity sites. Given the transcription-coupled recognition of CPDs, there is enhanced CPD repair in the template strand of actively transcribed genes within the gene body (Adar et al., 2016).

Several studies have focused on the influence of 3D genome organization on the formation and repair of double-strand breaks (DSBs). For instance, Sanders et al. (Sanders et al., 2020) demonstrated that post-irradiation 3D genome changes are cell type-specific, with an enhanced segregation of TADs noted in all tested repair-proficient cell types except for ATM-deficient fibroblasts. This indicates a potential mechanism to preserve 3D genome structure integrity during DNA damage repair.

Carré Simon et al. (Carre-Simon & Fabre, 2022) explored how chromatin functions within the DNA damage response to coordinate various cellular processes, including repair. They scrutinized the chromatin landscape before, during, and after DNA damage, with a focus on DSBs, and showed that chromatin modifications assist in the movement of both DSB-damaged and undamaged chromatin, thereby facilitating the mobilization, clustering, and repair of DSBs. Arnould et al. (Arnould, Rocher, Finoux, Clouaire, Li, Zhou, Caron, Mangeot, Ricci, Mourad, Haber, Noordermeer & Legube, 2021) revealed that TADs serve as functional units of the DNA damage response and are crucial for establishing γH2AX–53BP1 chromatin domains. They proposed a model whereby H2AX-containing nucleosomes are rapidly phosphorylated as they pass by DSB-anchored cohesin.

While these studies provide valuable insights, the relationship between the 3D structure of the genome and UV-induced damage formation and repair remains unexplored. Nonetheless, two previous studies have looked at the effect of 3D genome structure on UV-induced mutagenesis (García-Nieto, Schwartz, King, Paulsen, Collas, Herrera & Morrison, 2017; Perez, Wong, Schwartz, Herrera, King, García-Nieto & Morrison, 2021), suggesting that the outer regions of the genome are more susceptible to damage compared to the inner regions. In this study, we try to use the latest genome-wide mapping technologies for 3D genome, DNA damage, and repair, along with our in silico simulations, to uncover the interconnections between the 3D organization of the genome and DNA damage and repair.

2. READ SIMULATION WITH BOQUILA

The simulation of genomic data for the purpose of evaluating the performance of bioinformatics programs, particularly in the realms of read alignment, genome assembly, and variant and RNA-seq analysis, has gained considerable traction (Mangul, Martin, Hill, Lam, Distler, Zelikovsky, Eskin & Flint, 2019). This approach provides a structured means for performance evaluation even in situations where gold-standard data is unavailable. Notably, a majority of existing simulation tools are heavily inclined towards benchmarking; their focus is primarily on the generation of reads that emulate the output of a specific sequencing experiment by accurately mimicking the characteristics of the reads generated by sequencing machinery. Consequently, the metrics for correction predominantly pertain to artificial errors commonly introduced by these specific sequencing protocols.

Whilst the majority of these tools employ some form of simulation profile, they tend to replicate the characteristics of sequencing protocols, rather than biological experiments. For instance, the nucleotide content profile, which represents the proportion of each of the four nucleotides on a positional basis, is an element not considered in simulation tools, as per our knowledge. Numerous tools such as SomatoSim (Hawari, Hong & Biesecker, 2021), VarSim (Mu, Mohiyuddin, Li, Asadi, Gerstein, Abyzov, Wong & Lam, 2015), SimuSCoP (Yu, Du, Ban & Zhang, 2020), among others (Ivakhno, Colombo, Tanner, Tedder, Berri & Cox, 2017; Pattnaik, Gupta, Rao & Panda, 2014; Qin, Liu, Conroy, Morrison, Hu, Cheng, Murakami, Odunsi, Johnson, Wei, Liu & Wang, 2015; Xia, Liu, Deng & Xi, 2017; Yuan, Zhang & Yang, 2017), were specifically conceived for simulating genomic variation. Conversely, ART (Huang, Li, Myers & Marth, 2012) and SInC (Pattnaik et al., 2014) generate profiles based on error models and quality score distributions drawn from empirical data, whereas pIRS (Hu, Yuan, Shi, Lu, Liu, Li, Chen, Mu, Zhang, Li, Yue, Bai, Li & Fan, 2012) create quality profiles based on mapped reads and empirical data. Other tools such as NanoSim (Yang, Chu, Warren & Birol, 2017) and Gargammel, which simulate nanopore sequencing and ancient DNA sequencing respectively, use error

Results from this chapter have been published in (Akkose & Adebali, 2023a) profiles, length distributions, and can even mimic UV damage. However, these tools serve specific purposes and are not suitable for generating simulated datasets.

Notably, the nucleotide content of the reads can be biased for various reasons. Biases may be introduced during sequence library preparation that involves immunoprecipitation, ligation efficiency differences, or through the nature of the sequencing technology itself. For instance, sequencing methods that map UV damage typically result in dipyrimidine-enriched reads (Hu et al., 2016; Mao et al., 2016). Furthermore, the GC content of the reads, defined as the percentage of nitrogenous bases in a DNA or RNA molecule that are either guanine or cytosine, may vary depending on the sequencing platform (Ross, Russ, Costello, Hollinger, Lennon, Hegarty, Nusbaum & Jaffe, 2013). Lastly, the polymerase chain reaction (PCR) step might introduce another nucleotide bias due to differential efficiencies of universal primers towards specific nucleotides (Polz & Cavanaugh, 1998).

Given these factors that can impact the genomic distribution of reads, it is clear that there is a need for a sequencing read simulation tool that uses the nucleotide profile. While the above-mentioned simulation tools can account for the error and quality profiles of sequencing platforms and GC content biases, they are generally designed to simulate reads based on sequencing instruments and are not adequate for generating a simulated dataset that imitates the nucleotide content of input reads.

We, therefore, introduce boquila, a next-generation sequencing (NGS) read simulator that utilizes the nucleotide content profile. Boquila generates simulated reads that mirror the nucleotide profile of input reads, allowing the normalization of nucleotide content bias in actual reads by calculating the fold change between simulated and actual reads. Additionally, boquila is designed to utilize data from input sequencing when generating simulated reads, thereby enabling the use of these simulated reads to normalize the effects of copy number variations (CNV). Regions with a higher genomic copy number have an increased likelihood of being pulled down during library preparation, making those with lower copy numbers more challenging to detect. Thus, our approach provides an effective strategy for the generation of simulated datasets that authentically replicate the nucleotide content of input reads.

2.1 Generating Simulated Reads

Boquila was specifically designed to synthesize reads while mimicking the nucleotide composition of the input reads. It works with either FASTA or FASTQ files as input, and simulates reads based on the nucleotide content therein. The count and length distribution of the generated reads will mirror that of the input reads. However, reads containing ambiguous nucleotides (N) in the input data will be omitted from the simulation process. The nucleotide profile can be estimated based on either user-specified k-mer length or single nucleotides. Boquila can utilize either the entire genome or predefined genomic intervals, enabling random selection of reads from the reference genome and thus offering granular control over the regions where simulated reads are generated. Alternatively, if a user has access to raw genome sequencing data from a comparable experimental setup (same cell type, conditions, etc.), input DNA sequencing reads can be employed instead of the reference genome. When producing simulated reads, the nucleotide profile deduced from the input reads is dynamically adjusted according to the nucleotide profile obtained so far from the simulated reads. This ensures that the simulated reads closely resemble the input reads.

The software exports simulated reads in either FASTA or FASTQ format, based on the format of the input reads. It also provides an option for export in BED format. If input reads are available in FASTQ format, the quality scores are copied over to maintain equivalent mappability between the simulated and input reads. Alternatively, users can assign quality values to each synthesized read.

2.1.1 Read Simulation

Boquila initiates its process by calculating the nucleotide profile (NP) of the input reads. For each entry in the input data, 50 records are uniformly sampled from the reference genome. This number is adjustable and can be reduced to expedite read generation or increased to enhance the resemblance between the profiles of simulated and input reads, if necessary.

Subsequently, a score is assigned to each read, where each nucleotide is scored based on the frequency of its occurrence at the corresponding position in the input reads, using the NP of the input reads. The NP employed during read generation is adjusted at intervals - after every 10% of reads are generated. This adjustment is performed using the difference between the NP of input reads and the NP of reads simulated thus far. Half the difference between the NP of simulated reads at that stage and the NP of input reads (observed NP) is subtracted from the NP used for the simulation process, resulting in convergence towards the observed NP. This convergence process is performed after each decile of simulated reads, which are randomly selected from either the reference genome or input DNA sequencing data. Consequently, the order of input reads does not influence this convergence process.

2.1.2 Performance

In order to test the performance of Boquila, we conducted a test where reads for Escherichia coli XR-seq data (Adebali, Chiou, Hu, Sancar & Selby, 2017) were simulated. The test was executed on a compute cluster powered by an Intel Xeon Gold 6140 CPU @ 2.30GHz, running a Linux operating system. The test concluded in less than 14 minutes (Table 2.1), with the Fasta input format demonstrating a slightly faster processing time compared to Fastq. Additional tests were carried out to generate fixed-length reads (10bp and 20bp), where the runtime was found to increase linearly in correlation to the number of simulated reads. A doubling of the read length resulted in a 60% increase in runtime.

Input Format	Read length	Number of reads	Runtime (s)	Speed (no. of reads/s)
Fastq	Varied (17-31 bp)	15,279,119	838	18,232
Fasta	Varied (17-31 bp)	15,279,119	819	18,655
Fasta	10 bp	7,639,559	295	25,896
Fasta	10 bp	15,279,119	601	25,422
Fasta	20 bp	7,639,559	511	14,950
Fasta	20 bp	15,279,119	1049	14,565

Table 2.1 Boquila simulation performance

3. 3D GENOME ORGANIZATION AND UV-INDUCED DNA DAMAGE AND REPAIR

To better understand the impact of the genome's three-dimensional (3D) organization on the formation and subsequent repair of UV damage, we built a 3D model of the genome. The architecture of this model was based on Hi-C contact matrices and TADs derived from the HeLa cell line (Yardimci, Ozadam, Sauria, Ursu, Yan, Yang, Chakraborty, Kaul, Lajoie, Song, Zhan, Ay, Gerstein, Kundaje, Li, Taylor, Yue, Dekker & Noble, 2019). The architecture of each chromosome is depicted as a series of bead-like units, which represent TADs or inter-TAD regions. Each bead's size corresponds to the genomic area it represents (Figure 3.1A). Following the model construction, we divided the genome into discrete 1-μm slices, extending radially from the nucleus center to the periphery. This approach allowed us to analyze the distribution of UV damage and repair within these specific genomic areas. For the mapping of UV-induced damage sites and subsequent repair events, we utilized Damage-seq and XR-seq datasets (Huang, Azgari, Yin, Chiou, Lindsey-Boltz, Sancar, Hu & Adebali, 2022) respectively. These datasets, generated in HeLa cells, provide us with single nucleotide resolution of the damage and repair events. The damage distribution is determined immediately after the UV irradiation, precluding the possibility of repair at this early time point. We performed XR-seq 12 minutes post-UV irradiation, a time frame considered insufficient for the degradation of excised oligomers. Thus, we assumedly collected accumulated repair events at this specific 12-minute mark. Although this time point may raise expectations of observing the effects of transcription-coupled repair (TCR), particularly for cyclobutane pyrimidine dimer (CPD) repair, our previous research has demonstrated that TCR does not initiate at this juncture in HeLa cells (Huang et al., 2022). Therefore, our use of this dataset enabled us to exclude TCR and focus primarily on global repair events.

Results from this chapter have been published in (Akkose & Adebali, 2023b)

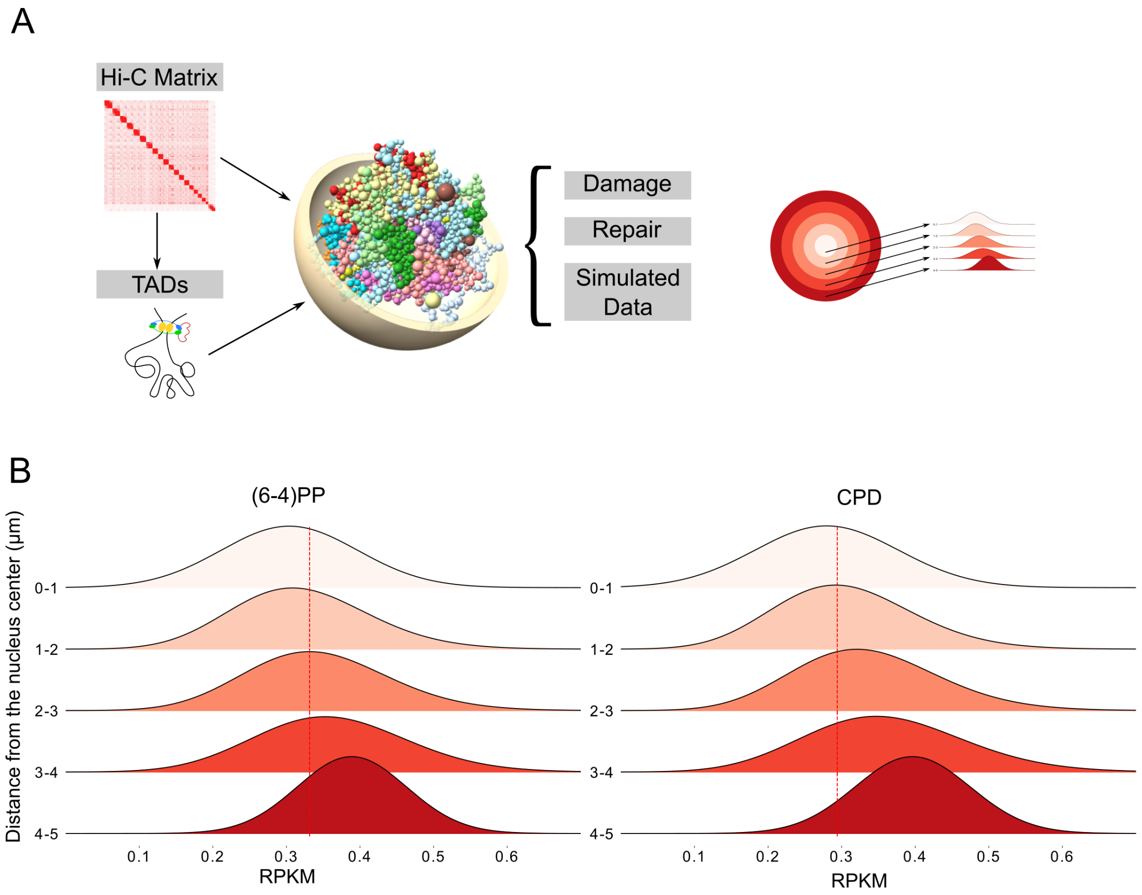

Figure 3.1 Study's methodology and the UV-induced damage distributions.

A, overall methodology of the study,one-micrometer nuclear sections shown on the right. B, the (6-4)PP and CPD damage data gathered right after UV exposure on 1-μm genomic sections are shown. The density of RPKM values for UV damage corresponding to each bead within the region is depicted. The dashed lines represent the median of the "0-1" region, defined as a sphere with a radius of 1 μm at the center of the nucleus. Welch's t-test was performed to compare the "0-1" region against all the remaining regions ("1-2", "2-3", "3-4", "4-5"). The resulting p-values for (6-4)PP were 0.123, 0.00052, 1.719e06, and 2.039e-11, while for CPD, they were 0.0307, 4.743e-06, 3.554e-10, and 2.639e-16 respectively. (6-4)PP, pyrimidine-pyrimidone (6-4); CPD cyclobutane pyrimidine dimer.

In our initial analysis, we mapped the distribution of UV-induced damage, collected instantaneously after UV irradiation (0 minutes), across the sliced sections of the 3D genome. We normalized the read counts by taking both the size of the region and the total number of mapped reads into account. Our findings revealed that in asynchronized HeLa cells, both types of UV-induced damage were prominently distributed in the genome's outermost regions, with a gradual decrease towards the nucleus's center (Figure 3.1B).

It is crucial to note that UV damage and repair maps are naturally skewed towards dipyrimidine-enriched sites (Hu et al., 2016; Mao et al., 2016). Cyclobutane pyrimidine dimers (CPDs) and pyrimidine-pyrimidone (6-4) photoproducts ((6-4)PPs) exhibit distinct nucleotide frequency profiles. Our initial goal was to analyze only the distribution of simulated reads, which reflects the genome's inherent nucleotide content bias in its 3D organization. We used our simulation tool, Boquila, to help with this. This tool randomly selects genomic regions from the reference genome or input DNA sequencing data such that the selected pseudo-reads exhibit a nucleotide frequency similar to the given NGS dataset (Akkose & Adebali, 2023a). It takes two inputs: (i) reference genome or pre-existing sequencing read data, and (ii) actual NGS data (in our case, XR-seq or Damage-seq). Boquila computes the nucleotide frequency distributions for the observed reads post-adapter trimming for each read length. Following this, it scans the reference sequence file and selects random reads. The tool employs a form of "closed-loop feedback" to continuously adapt the output to match the input read frequencies. It is worth noting that HeLa cells are a cancer cell line, which could potentially lead to chromosomal or regional copy number variations in our data, thereby influencing our results (Frattini, Fabbri, Valli, De Paoli, Montalbano, Gribaldo, Pasquali & Maserati, 2015). In this study, we used the HeLa input sequencing dataset as a reference to simulate the damage and repair reads (Huang et al., 2022). By generating simulated reads from input sequencing (low coverage whole-genome sequencing), the simulated data would include the effects of all copy number variations in the cell. Therefore, by using simulated data as a normalizing factor, we were able to eliminate the potential bias due to regional chromosomal copy number variation within the used HeLa cells when evaluating real damage and repair events, which would also be impacted by copy number variations. Read simulation utilizing input sequencing enabled us to correct genome-wide damage and repair distributions by eliminating chromosomal variations.

Following the simulation, we obtained randomly generated reads that collectively mimicked the nucleotide frequency distribution of the actual reads (Figure 3.2A). Surprisingly, when we mapped these simulated reads onto the 3D sections, we discovered that the concentration of (6-4)PP damage was higher in the genome's innermost regions. The frequency of these damages gradually decreased towards the nucleus periphery, which contrasts with the trend seen in the actual reads (Fig. 3B). CPD damage, on the other hand, was distributed more uniformly throughout the nucleus. To investigate whether this observation was HeLa-cell specific, we constructed 3D models for a variety of cell types: GM12878, KBM7, NHEK, HMEC, and HUVEC (Sanborn, Rao, Huang, Durand, Huntley, Jewett, Bochkov, Chinnappan, Cutkosky, Li, Geeting, Gnirke, Melnikov, McKenna, Stamenova, Lander & Aiden, 2015). We then examined the simulated UV damage distributions within the context of these 3D genomes. According to our findings, the simulated (6-4)PP and CPD damage distributions were consistently highest at the nucleus center and gradually decreased towards the outer regions for GM12878, KBM7, and NHEK cells, but not for the tested endothelial cells (Figure 3.2C). This suggests that the observed trend in simulated datasets is not exclusively linked to HeLa cells. However, while three of the five cell lines exhibited the same trend as HeLa cells, HMEC and HUVEC cell lines showed the same decreasing pattern except for the 0 to 1 μm region. The number of genomic regions falling into the innermost region in our 3D model was significantly fewer than other regions, a factor that might have introduced bias into our results.

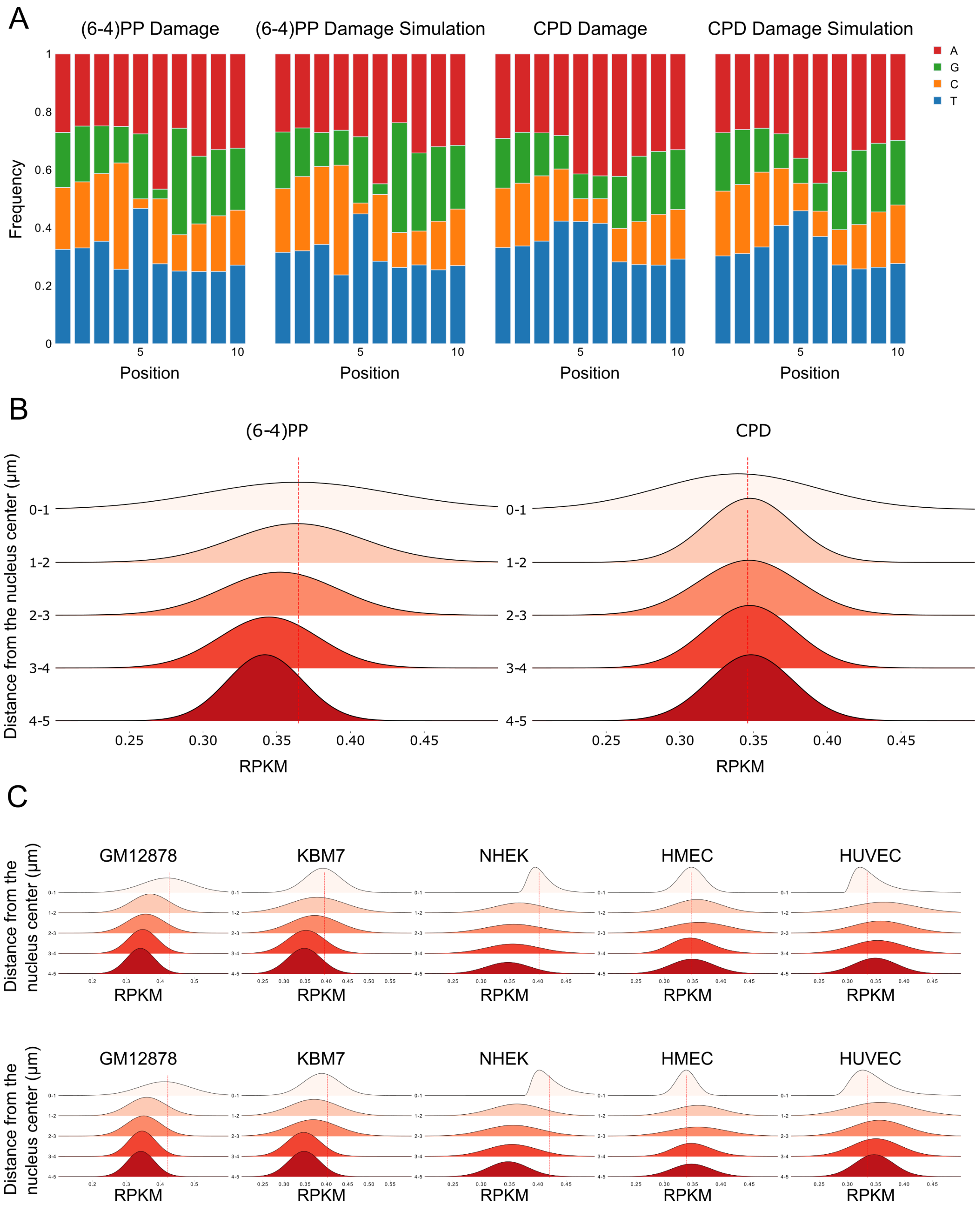

Figure 3.2 Simulated UV-induced damage sites for various cell lines

A, the nucleotide frequency for damage-seq (at 0 minutes) and the simulated damage-seq readings. Centralized damage-seq readings exhibit an increased presence of pyrimidines at the 5-6th positions. B, Expected (6-4)PP and CPD damage values (as per the simulation) on 1-μm genomic sections. The RPKM values for the simulated UV damage associated with each bead within the region were computed and the density of these RPKM values was shown. The dashed lines represent the median of the “0-1” region, which is a sphere with a 1-μm radius situated at the nucleus’s center. Welch’s t-test was performed to compare the “0-1” region against all the remaining regions (“1-2”, "2-3", "3-4", "4-5"). The derived p-values for (6-4)PP were 0.941, 0.997, 0.0097, and 0.0031, and for CPD, they were 0.201, 0.241, 0.203, and 0.16 respectively. C, simulated (6-4)PP (above) and CPD (below) damage values on the 3D genome models. These models were created using the Hi-C data from different cell lines. (6-4)PP, pyrimidine-pyrimidone (6-4); CPD cyclobutane pyrimidine dimer.

The simulated UV damage reads are randomly extracted from UV damage-prone sites, serving as an approximation of expected damage sites. To normalize the observed UV damage signal, we calculated the fold change between the actual and simulated damage signals, which we expressed as the observed-to-expected ratio. Following this normalization, the concentration of (6-4)PP and CPD damage (0 min) in asynchronized HeLa cells was found to be higher in the outermost regions of the genome, with a gradual decrease in frequency towards the nucleus's center (Figure 3.3). Interestingly, despite the presence of more UV-damaging regions in the inner parts of the genome, the concentration of UV damage was found to be greater in the outer regions. These findings indicate that the shielding effect of 3D genome organization for (6-4)PPs is greater than previously thought.

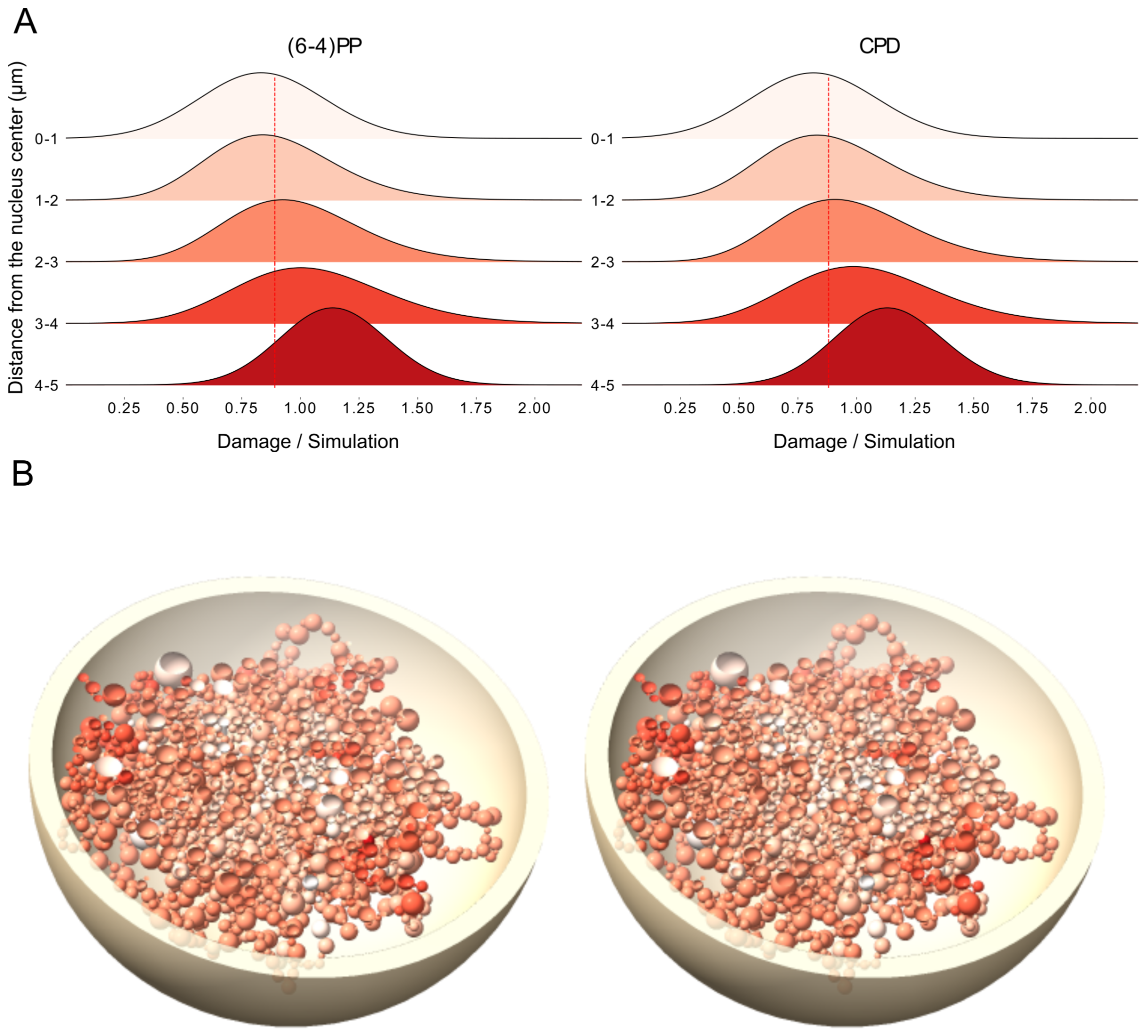

Figure 3.3 Normalized (observed over expected) damage formation for UV-induced damages.

A, the normalized damage values at 0 minutes for (6-4)PP and CPD in asynchronized cells on 1-μm genome sections. The RPKM value for each bead's UV damage was divided by the RPKM value of its simulated UV damage, and the density of these normalized damage values was shown. Dashed lines represent the median of the "0-1" region, described as a sphere with a 1-μm radius at the nucleus's center. Welch's t-test was performed to compare the "0-1" region against all the remaining regions ("1-2", "2-3", "3-4", "4-5"). The resulting p-values for (6-4)PP were 0.126, 2.55e-05, 1.46e-09, and 4.335e-15, while for CPD, they were 0.066, 9.651e-06, 9.7e-10, and 5.77e-16 respectively. B, a tomographic view of the data shown in A, with the (6-4)PP (left model) and CPD (right model) displaying beads in a gradient from white to red, representing increasing values. (6-4)PP, pyrimidine-pyrimidone (6-4); CPD cyclobutane pyrimidine dimer.

We used XR-seq data collected 12 minutes after UV irradiation in asynchronized HeLa cells to investigate the distributions of (6-4)PP and CPD repair within the 3D genome model. The initial analysis of non-normalized XR-seq data revealed no discernible differences between different 3D sections (Figure 3.4A). However, XR-seq reads, like damage reads, have a nucleotide content bias towards dipyrimidines due to damage sites. To address this inherent bias, we created simulated repair datasets and used them as a means of normalization. We began by looking at the distribution of repair levels (observed/expected) throughout the 3D layers. Surprisingly, we discovered that the outer regions produced more repair signals (Figure 3.4B). Yet, this is most likely due to greater damage formation on the periphery compared to the nucleus's center. As seen in Figure 3.1, the distribution of damage within the genome is not uniform, which may distort the genome-wide reported repair events obtained by XR-seq. To counteract the effect of the nonuniform initial damage formation observed at the zero minute mark, we calculated the fold change between the simulation-normalized repair levels and the simulation-normalized damage levels. When we examined these double-normalized repair values, we discovered that (6-4)PP and CPD repairs were evenly distributed across the genome (Figure 3.4C). There were no significant differences in normalized global excision repair 12 minutes after UV irradiation across the 3D layers.

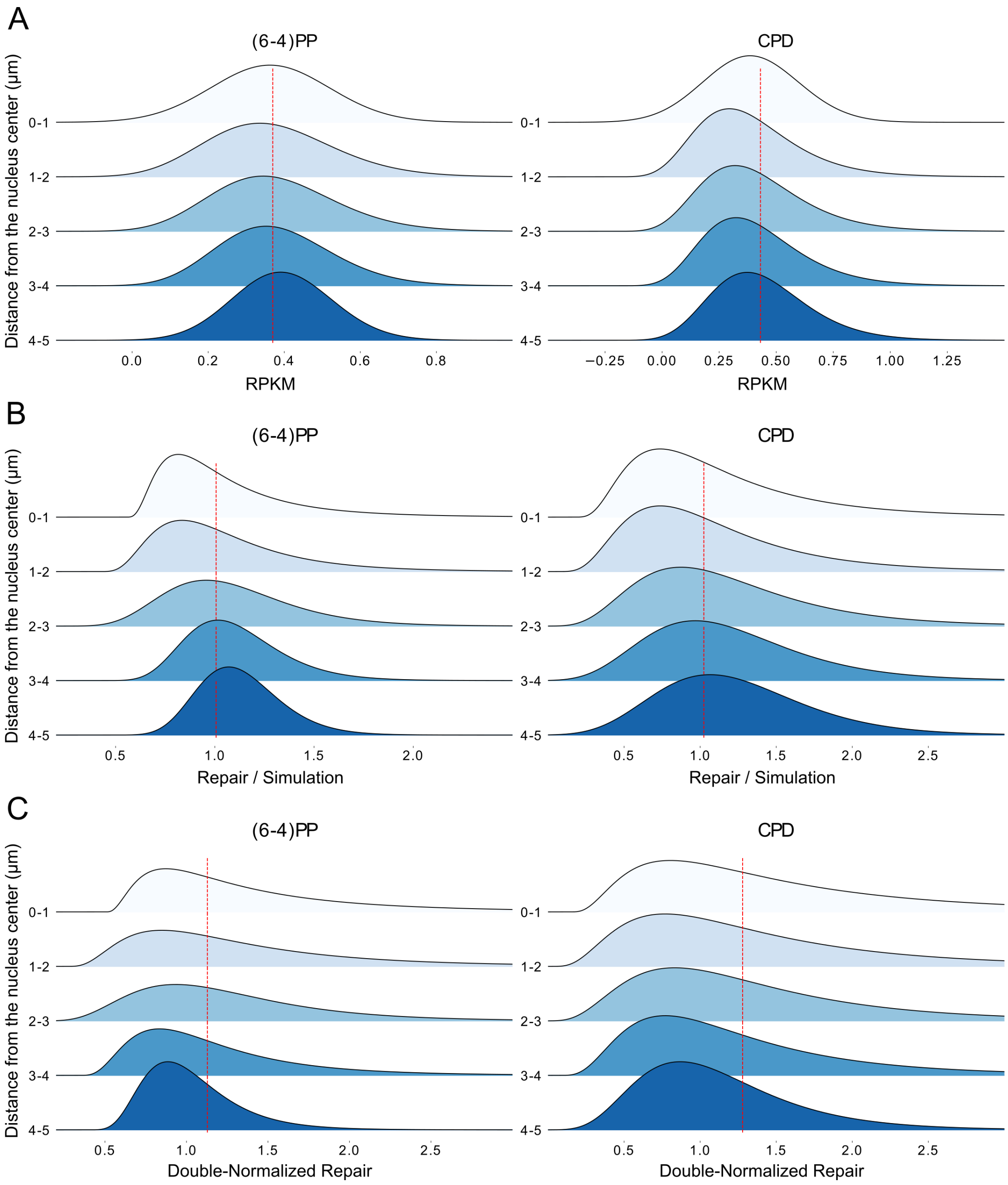

Figure 3.4 Global repair process within the 3D layers.

Repair 12 minutes post UV exposure (A), the normalized (Repair/Simulation) values (B), and the doubly normalized values (C) for (6-4)PP (left) and CPD (right). These repair values are based on 12-minute post-exposure repair, normalized by the damage observed at 0 minutes, with both further normalized by corresponding simulation data, presented on 1-μm genome slices. The density of these normalized repair values is shown. Dashed lines depict the median of the "0-1" region, described as a sphere with a 1-μm radius at the nucleus's center. Welch's t-test was performed to compare the “0-1” region against all the remaining regions (“1-2”, “2-3”, “3-4”, “4-5”). The p-values for (A) are 0.8, 0.458, 0.33, and 0.075 for (6-4)PP and 0.77, 0.49, 0.45, and

0.073 for CPD. For (B), the p-values are 0.846, 0.003, 3.57e-05, and 7.06e-07 for (6-4)PP and 0.6, 0.074, 0.034, and 0.0042 for CPD. For (C), the p-values are 0.91, 0.55, 0.214, and 0.05 for (6-4)PP and 0.46, 0.858, 0.162, and 0.05 for CPD. (6-4)PP, pyrimidine-pyrimidone (6-4); CPD cyclobutane pyrimidine dimer.

Transcription (Hu et al., 2017) and replication (Huang et al., 2022) are two integral cellular processes that may influence the formation and repair of DNA damage. Within the three-dimensional genomic landscape, gene transcription is not homogeneous; actively transcribed genes tend to be centrally located within the nucleus rather than at the periphery (Bickmore, 2013). Similarly, the spatial distribution of early and late replication domains within the 3D genome is not uniform. Early replication domains are predominantly located at the center of the genome, whereas late replication domains are more often found at the periphery (Sima, Chakraborty, Dileep, Michalski, Klein, Holcomb, Turner, Paulsen, Rivera-Mulia, Trevilla-Garcia, Bartlett, Zhao, Washburn, Nora, Kraft, Mundlos, Bruneau, Ljungman, Fraser, Ay & Gilbert, 2019). Therefore, we hypothesized that the observations made throughout our study could be influenced by these two important cellular processes rather than the spatial organization of the genome itself. To investigate this possibility, we performed the same analyses using 0-min Damage-seq and 12-min XR-seq data on both genic and intergenic regions (Figure A.1), as well as early and late replicating domains (Figure A.2). Our findings held true in both genic/intergenic areas and early/late replicating domains. This consistency shows that the effects we've seen are most likely due to the genome's three-dimensional conformational organization, rather than the influence of transcription or replication.

Our findings focused on the early repair events that occur in HeLa cells 12 minutes after UV irradiation, when only global repair is active. We examined CPD repair datasets collected two hours after UV irradiation to assess the possible impact of transcription-coupled repair (TCR). We used the same analytical technique as with the 12-minute samples to evaluate Damage-seq and XR-seq datasets taken two hours after UV irradiation from synchronized early-replicating and late-replicating HeLa cells. The 3D model was built using the same publicly available Hi-C dataset. In contrast to our 12-minute data, the two-hour samples revealed significant differences in repair level efficiency between the center and peripheral regions. In both early and late-replicating cells, central regions repaired more efficiently than their peripheral counterparts (Figure A.3). The 12-minute and two-hour samples differ primarily in two ways: (i) the extended time period following UV irradiation in the two-hour samples may allow for a reset of the genome's conformational organization, and (ii) the co-occurrence of TCR and global repair at two hours.

4. METHODS

4.1 Preprocessing Hi-C data and 3D genome modeling

Contact domains of the HeLa cell line were identified via analysis of Hi-C data, accessed under the Encode code ENCSR693GXU (Yardimci et al., 2019). Contact domains for GM12878, KBM7, NHEK, HMEC, and HUVEC cell lines were identified via analysis of Hi-C data obtained from GEO GSE63525 (Sanborn et al., 2015). 25 kb binned matrices were employed for TAD calling using the Arrowhead (Rao, Huntley, Durand, Stamenova, Bochkov, Robinson, Sanborn, Machol, Omer, Lander & Aiden, 2014) with its default parameters. Constructing models for the genomes of HeLa, GM12878, KBM7, NHEK, HMEC, and HUVEC was accomplished by utilizing the Chrom3D (Paulsen, Sekelja, Oldenburg, Barateau, Briand, Delbarre, Shah, Sorensen, Vigouroux, Buendia & Collas, 2017), adhering to the procedure described in the Paulsen et al. (Paulsen, Ali & Collas, 2018). To briefly outline the modeling process, overlapping TADs were integrated to formulate singular domains. For genomic regions not covered by a TAD, a proportionally sized bead was assigned. These bead dimensions were manipulated to constitute 15% of a modeled nucleus possessing a diameter of 10 μm. Interactions between these beads were deduced through an analysis of high-resolution Hi-C data, applicable to the respective cell lines.

Significant interactions were identified using a noncentral hypergeometric distribution(Paulsen et al., 2017) method articulated in Paulsen et al. (Paulsen et al., 2018). These interactions prompt bead pairs to gravitate towards one another, with the intention of reducing the spatial separation between them. A Monte Carlo optimization strategy was employed to minimize the bead-to-bead distances on a loss score function (Paulsen et al., 2017)

4.2 Damage-seq analysis

Unprocessed Damage-seq reads were obtained from SRA under the access code PRJNA608124 (Huang et al., 2022). Adapter sequences (GACTGGTTC-CAATTGAAAGTGCTCTTCCGATCT), were excised from the 5' ends of the raw Damage-seq reads employing the cutadapt (Martin, 2011). Subsequently, the trimmed reads were aligned to the Grch38 human genome utilizing bowtie2 (Langmead & Salzberg, 2012). Following alignment, the resulting BAM files were transformed into BED format using the bedtools suite (Quinlan & Hall, 2010). As the precise damage sites are located two nucleotides upstream of the reads, bedtools were employed to generate ten nucleotide-long reads, positioning the exact damage sites at the 5 and 6 nucleotide positions. Lastly, aligned reads were sorted and duplicate regions were removed.

4.3 XR-seq analysis

Unprocessed XR-seq reads were retrieved from the SRA under the access code PRJNA608124 (Huang et al., 2022). Subsequently, adapter sequences (TGGAATTCTCGGGTGCCAAGGAACTCCAGTNNNNNNACGATCTCGTAT-GCCGTCTTCTGCTTG) were excised from the 3' ends of these raw XR-seq reads using Cutadapt (Martin, 2011). Following the trimming process, trimmed reads were aligned to the Grch38 human genome utilizing bowtie2 (Langmead & Salzberg, 2012). Alignments in BAM format were then converted into BED format using the bedtools suite (Quinlan & Hall, 2010). Lastly, aligned reads were sorted and duplicate regions were removed.

4.4 Damage-seq and XR-seq simulations

Simulated datasets were produced using the software Boquila (v0.6) (Akkose & Adebali, 2023a). For the generation of simulated HeLa cells data, input DNA sequencing data were retrieved from SRA under the access code PRJNA608124. For other cell lines, GM12878, KBM7, NHEK, HMEC, and HUVEC, the hg19 human genome was used while generating the simulated reads.

When employing simulated data for the normalization of Damage-seq or XR-seq data, the simulation from the corresponding sample was utilized. As a consequence, for each true sample, we created simulated reads that were derived from these samples. The simulated reads were subsequently used to "correct" the corresponding damage or repair data.

4.5 Genic and intergenic regions

ENSEMBL genes were accessed from the BioMart (Smedley, Haider, Ballester, Holland, London, Thorisson & Kasprzyk, 2009). Using the bedtools suite (Quinlan & Hall, 2010), genes that overlapped were merged. Subsequently, these genes were intersected with beads in the 3D genomic models, which allowed us to identify genic regions. The regions that remained after this process were classified as intergenic regions.

4.6 Replication Domains

Replication domains were obtained from the SRA under the access code PRJNA608124 (Huang et al., 2022). The EdU-seq data was processed following the methodology detailed in the corresponding study (Huang et al., 2022). In summary, the read sequences were aligned to the GRCh38 human genome using the bowtie2 (Langmead & Salzberg, 2012). Samtools (Danecek, Bonfield, Liddle, Marshall, Ohan, Pollard, Whitwham, Keane, McCarthy, Davies & Li, 2021) was employed to eliminate reads of quality less than 20 and duplicate reads. Following this, the ratio of early to late reads was computed in 50 kb long windows and log2-transformed. Finally, replication domains were created using a custom R script (Huang et al., 2022).

5. DISCUSSION

In this study, Using Damage-seq and Hi-C-seq tests using the same cell line, we evaluated and confirmed the hypothesis proposing a shielding effect (García-Nieto et al., 2017) of the three-dimensional structure of nuclear DNA against UV irradiation. Our findings imply that this protective impact is significantly stronger than previously thought. Potential damage sites are significantly less common on the nuclear periphery than in the center. These possible damage sites were discovered using computational simulations, and the observed and expected damage events were compared, highlighting the importance of the shielding effect.computational simulations, and their observed and expected damage events were compared, thereby reinforcing the prominence of the shielding effect.

When compared to actual reads, our efforts to replicate read data resulted in similar nucleotide frequency distributions. This simulated data helps to answer a fundamental question: how many occurrences of damage would we predict inside a certain region in the absence of a substantial genetic component (such as 3D organization) driving damage formation? The simulated reads created a "expected" damage count based solely on nucleotide content by mimicking the nucleotide content of the actual reads. We were able to distinguish the influence of genomic variables independent of nucleotide content by comparing observed to expected damage.

One unanswered topic is why potential UV-induced damage locations differ between the nucleus's center and peripherial areas. Surprisingly, this differential effect is only seen in the nucleotide profiles of (6-4)PPs and not in CPDs. The difference between these two types of UV-induced damage is due to the presence of thymine-cytosine (TC) sites, which are more prevalent in (6-4)PPs (Hu et al., 2015). As a result, this difference is due to TCs in (6-4)PPs. Despite being less frequent than CPDs, (6-4)PPs are typically more mutagenic (LeClerc, Borden & Lawrence, 1991). Furthermore, due to the unique chemistry of the dipyrimidine bulky adduct, these two types of damage result in different structures of helix distortion (Kim, Patel & Choi, 1995). The global nucleotide excision repair machinery recognizes (6-4)PP damage faster than CPD damage. (Hu, Choi, Gaddameedhi, Kemp, Reardon &

Sancar, 2013; Mu, Tursun, Duckett, Drummond, Modrich & Sancar, 1997). CPD repair, on the other hand, is more prone to transcription-coupled repair (TCR) because global repair mechanisms are insufficient for the timely recognition of these bulky adducts (Hu et al., 2017). Given these factors, despite their lower frequency, (6-4)PPs pose a greater threat to genome integrity than CPDs. The threat posed by (6-4)PPs may have influenced the evolution of 3D genomic organization. Peripheral TC sites, which experience damage more frequently than core sites, may have been exposed to UV radiation over time, resulting in TC conversion to TT via C>T mutation. This repeated substitution could lead to fewer TC sites in peripheral regions, resulting in the current 3D organization with a lower risk of (6-4)PP damage at the periphery. From an evolutionary standpoint, this adaptation may benefit genomic integrity by lowering the "dangerous" regions prone to UV damage near the periphery.

The universality of the observed trend toward fewer possible UV damage sites at the nuclear perimeter remains an open question. Although four of the six cell lines evaluated show rising levels of possible UV damage sites or dipyrimidines, the remaining two cell lines deviate from this trend between 0 and 1 *mu*m. The trend is visible for the remaining sections from 1 to 5 *mu*m. This disparity could be attributed to the minimal number of genetic regions contained inside this innermost area. Or, these two cell lines may have differentiated, resulting in a higher concentration of active genes essential for their unique expression patterns within the innermost areas. These genes may have a nucleotide bias that results in a higher number of pyrimidine sites.

In addition, we explored the impact of 3D genome organization on nucleotide excision repair. Given that peripheral regions of the nucleus are predominantly composed of heterochromatic regions (Bickmore, 2013), and such regions have been reported to exhibit poor repair characteristics (Adar et al., 2016) we hypothesized that the core regions would exhibit preferential repair relative to peripheral regions. Further, the peripheral regions consist largely of late-replicating domains, while central regions are mainly composed of early-replicating domains. Our previous work demonstrated that early-replicating domains are more efficiently repaired than late-replicating domains (Huang et al., 2022). Accordingly, we anticipated the peripheral regions to be less efficiently repaired than the core regions. However, contrary to our expectations, we did not observe a significant difference in repair efficiency across different sections of the spherical genomes in samples collected 12 minutes post-UV irradiation. This finding suggests that UV irradiation may alter the 3D organization of the genome, affecting subsequent repair efficiency. It's worth noting that, although no direct evidence currently supports the notion that UV irradiation induces changes in the 3D organization of the genome, previous studies have shown that gamma irradiation can alter the 3D genome structure (Sanders et al., 2020). Consequently, it's plausible that by the time we assess the repair events 12 minutes post-irradiation, the 3D organization of the genome might have undergone changes. Nevertheless, these potential alterations should not have affected damage formation since the damage data were collected immediately following UV irradiation. To summarize, while it's probable that repair processes are more efficient in the central regions of the nucleus, we were unable to ascertain this trend due to the potential alterations in the 3D organization of nuclear DNA 12 minutes after UV irradiation. Another potential explanation for this could be the presence of robust global repair mechanisms that efficiently mitigate the damage throughout the genome. We have previously demonstrated that 12 minutes following UV irradiation, global repair is sufficiently active, rendering the need for transcription-coupled repair (TCR) to recognize extensive damage unnecessary. Therefore, at this time point, we did not observe the preferred repair of the transcribed strand for either CPDs or (6-4)PPs in HeLa cells. Given that TCR is more affected by chromatin structure than global repair, the absence of observable differences in repair levels may, in part, be attributed to the high efficiency of genome-wide global repair.

Despite the fact that we did not observe differential repair levels across 3D sections of the genome after 12 minutes, we did observe more efficient repair of CPDs in HeLa cells two hours after UV irradiation, while both global repair and TCR are operational. The observed repair efficiency within the nucleus's center region was unaffected by the cell's replication phase; both early and late-replicating synchronized cells had similar profiles. This finding suggests that chromatin has a greater influence on TCR than global repair. Furthermore, it is reasonable to expect a possible recovery of the 3D genome organization after 2 hours of UV irradiation.

In conclusion, we have not only supported but also expanded on the idea of DNA shielding using Damage-seq, simulated Damage-seq, and Hi-C-seq datasets. We discovered diverse patterns of potential damage locations within different parts of the 3D genomic organization for (6-4)PPs but not CPDs. The absence of such a pattern in CPDs shows that there is an evolutionary pressure to decrease theoretical (6-4)PP locations on the periphery. The lack of discernible changes in repair efficiency between the core and periphery 12 minutes after UV irradiation suggests that UV irradiation changes the genome's 3D conformation.

BIBLIOGRAPHY

Adar, S., Hu, J. C., Lieb, J. D., & Sancar, A. (2016). Genome-wide kinetics of dna excision repair in relation to chromatin state and mutagenesis. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(15), E2124–E2133.

Adebali, O., Chiou, Y. Y., Hu, J., Sancar, A., & Selby, C. P. (2017). Genome-wide transcription-coupled repair in escherichia coli is mediated by the mfd translocase. *Proc Natl Acad Sci U S A*, *114*(11), E2116–E2125.

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

Akkose, U. & Adebali, O. (2023b). The interplay of 3d genome organization with UV-induced DNA damage and repair. *Journal of Biological Chemistry*, *299*(5), 104679.

Arnould, C., Rocher, V., Finoux, A. L., Clouaire, T., Li, K., Zhou, F. L., Caron, P., Mangeot, P. E., Ricci, E. P., Mourad, R., Haber, J. E., Noordermeer, D., & Legube, G. (2021). Loop extrusion as a mechanism for formation of dna damage repair foci. *Nature*, *590*(7847).

Bickmore, W. A. (2013). The spatial organization of the human genome. *Annu Rev Genomics Hum Genet*, *14*, 67–84.

Carre-Simon, A. & Fabre, E. (2022). 3d genome organization: Causes and consequences for dna damage and repair. *Genes*, *13*(1).

Chakraborty, A. & Ay, F. (2019). The role of 3d genome organization in disease: From compartments to single nucleotides. *Semin Cell Dev Biol*, *90*, 104–113.

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of samtools and bcftools. *Gigascience*, *10*(2).

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), 376–80.

Frattini, A., Fabbri, M., Valli, R., De Paoli, E., Montalbano, G., Gribaldo, L., Pasquali, F., & Maserati, E. (2015). High variability of genomic instability and gene expression profiling in different hela clones. *Sci Rep*, *5*, 15377.

García-Nieto, P. E., Schwartz, E. K., King, D. A., Paulsen, J., Collas, P., Herrera, R. E., & Morrison, A. J. (2017). Carcinogen susceptibility is regulated by genome architecture and predicts cancer mutagenesis. *The EMBO Journal*, *36*(19), 2829–2843.

Hawari, M. A., Hong, C. S., & Biesecker, L. G. (2021). Somatosim: precision simulation of somatic single nucleotide variants. *Bmc Bioinformatics*, *22*(1).

Hou, C. H., Li, L., Qin, Z. H. S., & Corces, V. G. (2012). Gene density, transcription, and insulators contribute to the partition of the drosophila genome into physical domains. *Molecular Cell*, *48*(3), 471–484.

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

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

Hu, J. C., Choi, J. H., Gaddameedhi, S., Kemp, M. G., Reardon, J. T., & Sancar, A. (2013). Nucleotide excision repair in human cells fate of the excised oligonucleotide carrying dna damage in vivo. *Journal of Biological Chemistry*, *288*(29), 20918–20926.

Hu, J. C., 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 of the United States of America*, *113*(41), 11507–11512.

Hu, X. S., Yuan, J. Y., Shi, Y. J., Lu, J. L., Liu, B. H., Li, Z. Y., Chen, Y. X., Mu, D. S., Zhang, H., Li, N., Yue, Z., Bai, F., Li, H., & Fan, W. (2012). pirs: Profile-based illumina pair-end reads simulator. *Bioinformatics*, *28*(11), 1533–1535.

Huang, W. C., Li, L. P., Myers, J. R., & Marth, G. T. (2012). Art: a next-generation sequencing read simulator. *Bioinformatics*, *28*(4), 593–594.

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

Ivakhno, S., Colombo, C., Tanner, S., Tedder, P., Berri, S., & Cox, A. J. (2017). thapmix: simulating tumour samples through haplotype mixtures. *Bioinformatics*, *33*(2), 280–282.

Kim, J. K., Patel, D., & Choi, B. S. (1995). Contrasting structural impacts induced by cis-syn cyclobutane dimer and (6-4)-adduct in dna duplex decamers - implication ln mutagenesis and repair activity. *Photochemistry and Photobiology*, *62*(1), 44–50.

Langmead, B. & Salzberg, S. L. (2012). Fast gapped-read alignment with bowtie 2. *Nature Methods*, *9*(4), 357–U54.

LeClerc, J. E., Borden, A., & Lawrence, C. W. (1991). The thymine-thymine pyrimidine-pyrimidone(6-4) ultraviolet light photoproduct is highly mutagenic and specifically induces 3' thymine-to-cytosine transitions in escherichia coli. *Proceedings of the National Academy of Sciences*, *88*(21), 9685–9689.

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

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., & Dekker, J. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, *326*(5950), 289–93.

Mangul, S., Martin, L. S., Hill, B. L., Lam, A. K. M., Distler, M. G., Zelikovsky, A., Eskin, E., & Flint, J. (2019). Systematic benchmarking of omics computational tools. *Nature Communications*, *10*.

Mao, P., Smerdon, M. J., Roberts, S. A., & Wyrick, J. J. (2016). Chromosomal landscape of uv damage formation and repair at single-nucleotide resolution. *Proceedings of the National Academy of Sciences of the United States of Amer-*

ica, *113*(32), 9057–9062.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal*.

Mu, D., Tursun, M., Duckett, D. R., Drummond, J. T., Modrich, P., & Sancar, A. (1997). Recognition and repair of compound dna lesions (base damage and mismatch) by human mismatch repair and excision repair systems. *Molecular and Cellular Biology*, *17*(2), 760–769.

Mu, J. C., Mohiyuddin, M., Li, J., Asadi, N. B., Gerstein, M. B., Abyzov, A., Wong, W. H., & Lam, H. Y. K. (2015). Varsim: a high-fidelity simulation and validation framework for high-throughput genome sequencing with cancer applications. *Bioinformatics*, *31*(9), 1469–1471.

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N. L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Bluthgen, N., Dekker, J., & Heard, E. (2012). Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, *485*(7398), 381–385.

Pattnaik, S., Gupta, S., Rao, A. A., & Panda, B. (2014). Sinc: an accurate and fast error-model based simulator for snps, indels and cnvs coupled with a read generator for short-read sequence data. *Bmc Bioinformatics*, *15*.

Paulsen, J., Ali, T. M. L., & Collas, P. (2018). Computational 3d genome modeling using chrom3d. *Nature Protocols*, *13*(5), 1137–1152.

Paulsen, J., Sekelja, M., Oldenburg, A. R., Barateau, A., Briand, N., Delbarre, E., Shah, A., Sorensen, A. L., Vigouroux, C., Buendia, B., & Collas, P. (2017). Chrom3d: three-dimensional genome modeling from hi-c and nuclear lamin-genome contacts. *Genome Biology*, *18*.

Perez, B. S., Wong, K. M., Schwartz, E. K., Herrera, R. E., King, D. A., García-Nieto, P. E., & Morrison, A. J. (2021). Genome-wide profiles of uv lesion susceptibility, repair, and mutagenic potential in melanoma. *Mutat Res*, *823*, 111758.

Polz, M. F. & Cavanaugh, C. M. (1998). Bias in template-to-product ratios in multitemplate pcr. *Applied and Environmental Microbiology*, *64*(10), 3724–3730.

Pope, B. D., Ryba, T., Dileep, V., Yue, F., Wu, W., Denas, O., Vera, D. L., Wang, Y., Hansen, R. S., Canfield, T. K., Thurman, R. E., Cheng, Y., Gulsoy, G., Dennis, J. H., Snyder, M. P., Stamatoyannopoulos, J. A., Taylor, J., Hardison, R. C., Kahveci, T., Ren, B., & Gilbert, D. M. (2014). Topologically associating domains are stable units of replication-timing regulation. *Nature*, *515*(7527), 402–5.

Qin, M. C., Liu, B., Conroy, J. M., Morrison, C. D., Hu, Q., Cheng, Y. B., Murakami, M., Odunsi, A. O., Johnson, C. S., Wei, L., Liu, S., & Wang, J. M. (2015). Scnvsim: somatic copy number variation and structure variation simulator. *Bmc Bioinformatics*, *16*.

Quinlan, A. R. & Hall, I. M. (2010). Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842.

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., Sanborn, A. L., Machol, I., Omer, A. D., Lander, E. S., & Aiden, E. L. (2014). A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, *159*(7), 1665–1680.

Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C., & Jaffe, D. B. (2013). Characterizing and measuring bias in sequence data. *Genome Biology*, *14*(5).

Sanborn, A. L., Rao, S. S., Huang, S. C., Durand, N. C., Huntley, M. H., Jewett, A. I., Bochkov, I. D., Chinnappan, D., Cutkosky, A., Li, J., Geeting, K. P., Gnirke, A., Melnikov, A., McKenna, D., Stamenova, E. K., Lander, E. S., & Aiden, E. L. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc Natl Acad Sci U S A*, *112*(47), E6456–65.

Sanders, J. T., Freeman, T. F., Xu, Y., Golloshi, R., Stallard, M. A., Hill, A. M., San Martin, R., Balajee, A. S., & McCord, R. P. (2020). Radiation-induced dna damage and repair effects on 3d genome organization. *Nat Commun*, *11*(1), 6178.

Sanyal, A., Lajoie, B. R., Jain, G., & Dekker, J. (2012). The long-range interaction landscape of gene promoters. *Nature*, *489*(7414), 109–U127.

Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N. A., Huber, W., Haering, C. H., Mirny, L., & Spitz, F. (2017). Two independent modes of chromatin organization revealed by cohesin removal. *Nature*, *551*(7678), 51–56.

Sima, J., Chakraborty, A., Dileep, V., Michalski, M., Klein, K. N., Holcomb, N. P., Turner, J. L., Paulsen, M. T., Rivera-Mulia, J. C., Trevilla-Garcia, C., Bartlett, D. A., Zhao, P. A., Washburn, B. K., Nora, E. P., Kraft, K., Mundlos, S., Bruneau, B. G., Ljungman, M., Fraser, P., Ay, F., & Gilbert, D. M. (2019). Identifying cis elements for spatiotemporal control of mammalian dna replication. *Cell*, *176*(4), 816–830 e18.

Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., & Kasprzyk, A. (2009). Biomart - biological queries made easy. *Bmc Genomics*, *10*.

Woodcock, C. L. & Dimitrov, S. (2001). Higher-order structure of chromatin and chromosomes. *Curr Opin Genet Dev*, *11*(2), 130–5.

Xia, Y. C., Liu, Y., Deng, M. H., & Xi, R. B. (2017). Pysim-sv: a package for simulating structural variation data with gc-biases. *Bmc Bioinformatics*, *18*.

Yang, C., Chu, J., Warren, R. L., & Birol, I. (2017). Nanosim: nanopore sequence read simulator based on statistical characterization. *Gigascience*, *6*(4).

Yardimci, G. G., Ozadam, H., Sauria, M. E. G., Ursu, O., Yan, K. K., Yang, T., Chakraborty, A., Kaul, A., Lajoie, B. R., Song, F., Zhan, Y., Ay, F., Gerstein, M., Kundaje, A., Li, Q. H., Taylor, J., Yue, F., Dekker, J., & Noble, W. S. (2019). Measuring the reproducibility and quality of hi-c data. *Genome Biology*, *20*.

Yu, Z. H., Du, F., Ban, R. J., & Zhang, Y. W. (2020). Simuscop: reliably simulate illumina sequencing data based on position and context dependent profiles. *Bmc Bioinformatics*, *21*(1).

Yuan, X. G., Zhang, J. Y., & Yang, L. Y. (2017). Intsim: An integrated simulator of next-generation sequencing data. *Ieee Transactions on Biomedical Engineering*, *64*(2), 441–451.

APPENDIX A

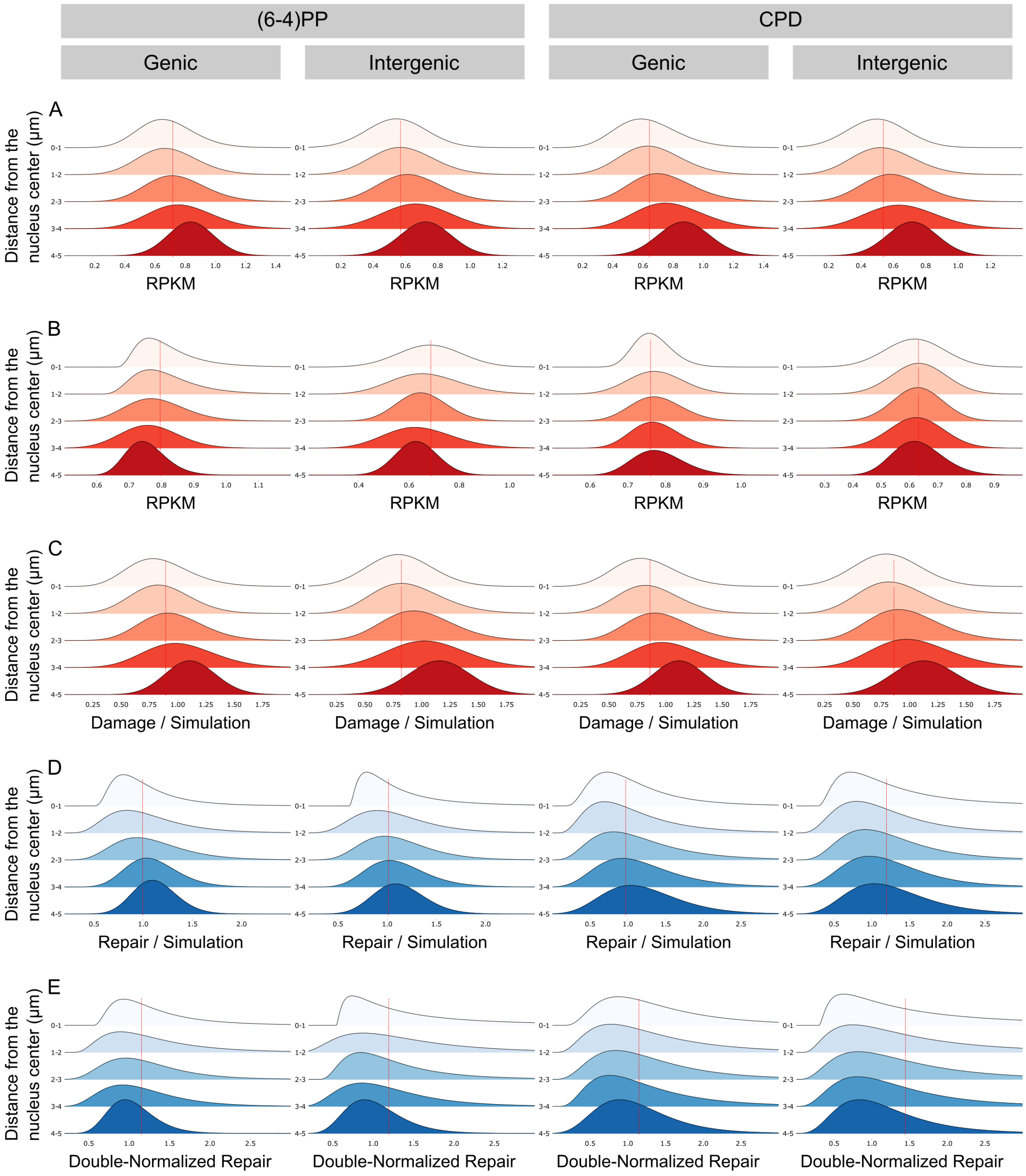

Figure A.1 Damage and repair in genic and intergenic domains.

A: (6-4)PP and CPD damage values collected immediately after UV irradiation for genic and intergenic regions on one-micrometer genome slices. RPKM values of UV damage for each bead in the region were calculated and the density of RPKM values of the beads was shown. Dashed lines depict the median of the "0-1" region, described as a sphere with a 1-μm radius at the nucleus's center. Welch's t-test was performed to compare the “0-1” region against all the remaining regions (“1-2”, “2-3”, “3-4”, “4-5”), p-values are: 0.59, 0.0036, 2.39e-05, 5.82e-10 for Genic-(6-4)PP; 0.104, 0.0002, 1.13e-06, 1.49e-11 for Intergenic-(6-4)PP; 0.0307, 0.0001, 3.7e-8, 1.29e-14 for Genic-CPD and 0.033, 1.12e-05, 1.18e-09, 4.214e-15 for Intergenic-CPD, respectively.

B: (6-4)PP and CPD simulated damage values (based on 0 min Damage-seq) for genic and intergenic regions on one-micrometer genome slices. RPKM values of simulated UV damage for each bead in the region are calculated and the density of RPKM values of the beads was shown. Dashed lines depict the median of the "0-1" region, described as a sphere with a 1-μm radius at the nucleus's center. Welch's t-test was performed to compare the "0-1" region against all the remaining regions ("1-2", "2-3", "3-4", "4-5"), p-values are: 0.75, 0.01, 8.6e-05, 6.02e-06 for Genic-(6-4)PP; 0.51, 0.051, 0.024, 0.0016, 0.008 for Intergenic-(6-4)PP; 0.32, 0.163, 0.254, 0.029 for Genic-CPD and 0.245, 0.366, 0.649, 0.273 for Intergenic-CPD, respectively.

C: (6-4)PP and CPD normalized damage values (0 min) for genic and intergenic regions on one-micrometer genome slices. RPKM value of UV damage for each bead divided by the RPKM value of simulated UV damage for it and the density of normalized damage values of the beads was shown. Dashed lines depict the median of the "0-1" region, described as a sphere with a 1-μm radius at the nucleus's center. Welch's t-test was performed to compare the "0-1" region against all the remaining regions ("1-2", "2-3", "3-4", "4-5"), p-values are: 0.7, 0.0009, 4.17e-07, 7.9e-12 for Genic-(6-4)PP; 0.03, 5.46e-06, 1.27e-09, 1.87e-15 for Intergenic-(6-4)PP; 0.4, 0.00035, 1.8e-07, 6.67e-13 for Genic-CPD and 0.082, 3.63e-05, 6.01e-09, 2.71e-14 for Intergenic-CPD, respectively.

D: Normalized (Repair / Simulation) repair values (XR-seq collected 12 min after UV) for genic and intergenic regions on 1 micrometer genome slices. Density of normalized repair values of the beads were shown.Dashed lines depict the median of the "0-1" region, described as a sphere with a 1-μm radius at the nucleus's center. Welch's t-test was performed to compare the "0-1" region against all the remaining regions ("1-2", "2-3", "3-4", "4-5"), p-values are: 0.82, 0.0034, 1.09e-05, 4.25e-07 for Genic-(6-4)PP; 0.063, 0.078, 0.007, 0.0009 for Intergenic-(6-4)PP; 0.78, 0.07, 0.038, 0.0028 for Genic-CPD and 0.29, 0.635, 0.415, 0.21 for Intergenic-CPD, respectively.

E: Double-normalized repair values (XR-seq collected 12 min, Damage-seq collected 0 min after UV) for genic and intergenic regions on 1 micrometer genome slices. Density of normalized repair values of the beads were shown. Dashed lines depict the median of the "0-1" region, described as a sphere with a 1-μm radius at the nucleus's center. Welch's t-test was performed to compare the "0-1" region against all the remaining regions ("1-2", "2-3", "3-4", "4-5"), p-values are: 0.63, 0.436, 0.28, 0.042 for Genic-(6-4)PP; 0.3, 0.24, 0.19, 0.12 for Intergenic-(6-4)PP; 0.67, 0.37, 0.24, 0.043 for Genic-CPD and 0.28, 0.25, 0.18, 0.13 for Intergenic-CPD, respectively.

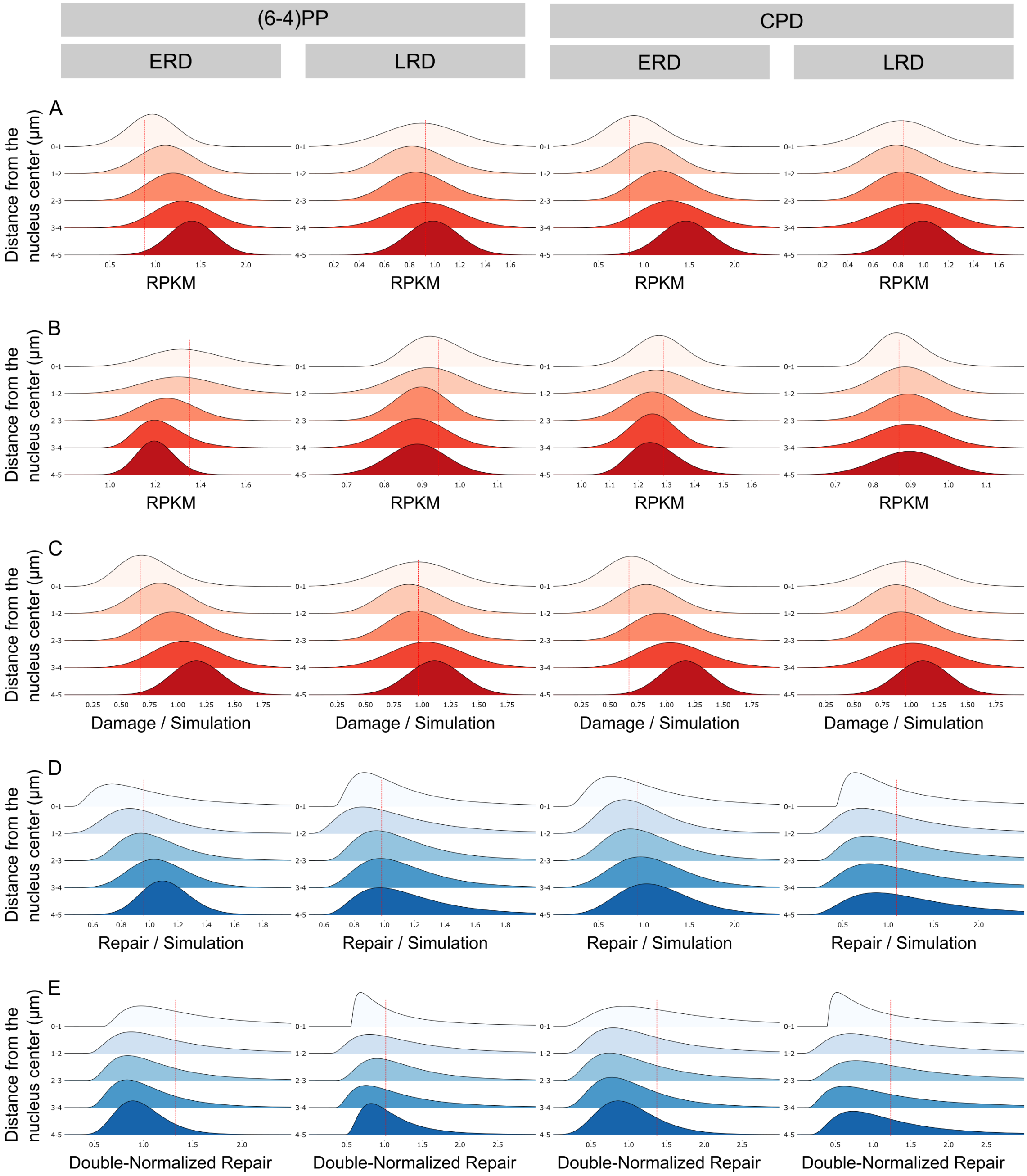

Figure A.2 Damage and repair in early/late replicating domains.

A: (6-4)PP and CPD damage values collected immediately after UV irradiation for early replication domains (ERD) and late replication domains (LRD) on one-micrometer genome slices. RPKM values of UV damage for each bead in the region were calculated and the density of RPKM values of the beads was shown. Dashed lines depict the median of the "0-1" region, described as a sphere with a 1-μm radius at the nucleus's center. Welch's t-test was performed to compare the "0-1" region against all the remaining regions ("1-2", "2-3", "3-4", "4-5"), p-values are: 0.001, 4.46e-08, 1.7e-11, 1.17e-14 for ERD-(6-4)PP; 0.045, 0.76, 0.53, 0.1 for LRD-(6-4)PP; 0.0028, 2.8e-08, 3.27e-12, 2.11e-16 for ERD-CPD and 0.94, 0.32, 0.025, 0.0009 for LRD-CPD, respectively.

B: (6-4)PP and CPD simulated damage values (based on 0 min Damage-seq) for early replication domains (ERD) and late replication domains (LRD) on one-micrometer genome slices. RPKM values of simulated UV damage for each bead in the region are calculated and the density of RPKM values of the beads was shown. Dashed lines depict the median of the "0-1" region, described as a sphere with a 1-μm radius at the nucleus's center. Welch's t-test was performed to compare the “0-1” region against all the remaining regions (“1-2”, “2-3”, “3-4”, “4-5”), p-values are: 0.34, 0.001, 4.9e-05, 1.58e-06 for ERD-(6-4)PP; 0.07, 0.0025, 0.00017, 0.0004 for LRD-(6-4)PP; 0.45, 0.094, 0.091, 0.39 for ERD-CPD and 0.56, 0.2, 0.24, 0.16 for LRD-CPD, respectively.

C: (6-4)PP and CPD normalized damage values (0 min) for early replication domains (ERD) and late replication domains (LRD) on one-micrometer genome slices. RPKM value of UV damage for each bead divided by the RPKM value of simulated UV damage for it and the density of normalized damage values of the beads was shown. Dashed lines depict the median of the "0-1" region, described as a sphere with a 1-μm radius at the nucleus's center. Welch's t-test was performed to compare the “0-1” region against all the remaining regions (“1-2”, “2-3”, “3-4”, “4-5”), p-values are: 0.0048, 2.77e-08, 3.66e-12, 1.2e-15 for ERD-(6-4)PP; 0.74, 0.68, 0.12, 0.001 for LRD-(6-4)PP; 0.0018, 5.03e-09, 6.11e-13, 1.65e-16 for ERD-CPD and 0.9, 0.54, 0.072, 0.0047 for LRD-CPD, respectively.

D: Normalized (Repair / Simulation) repair values (XR-seq collected 12 min after UV) for early replication domains (ERD) and late replication domains (LRD) on one-micrometer genome slices. Density of normalized repair values of the beads were shown. Dashed lines depict the median of the "0-1" region, described as a sphere with a 1-μm radius at the nucleus's center. Welch's t-test was performed to compare the “0-1” region against all the remaining regions (“1-2”, “2-3”, “3-4”, “4-5”), p-values are: 0.6, 0.0077, 0.00013, 6.25e-06 for ERD-(6-4)PP; 0.45, 0.258, 0.2, 0.13 for LRD-(6-4)PP; 0.89, 0.164, 0.043, 0.0057 for ERD-CPD and 0.445, 0.73, 0.86, 0.737 for LRD-CPD, respectively.

E: Double-normalized repair values (XR-seq collected 12 min, Damage-seq collected 0 min after UV) for early replication domains (ERD) and late replication domains (LRD) on one-micrometer genome slices. Density of normalized repair values of the beads were shown. Dashed lines depict the median of the "0-1" region, described as a sphere with a 1-μm radius at the nucleus's center. Welch's t-test was performed to compare the “0-1” region against all the remaining regions (“1-2”, “2-3”, “3-4”, “4-5”), p-values are: 0.46, 0.37, 0.11, 0.06 for ERD-(6-4)PP; 0.69, 0.47, 0.87, 0.37 for LRD-(6-4)PP; 0.28, 0.79, 0.08, 0.046 for ERD-CPD and 0.53, 0.45, 0.94, 0.29 for LRD-CPD, respectively.

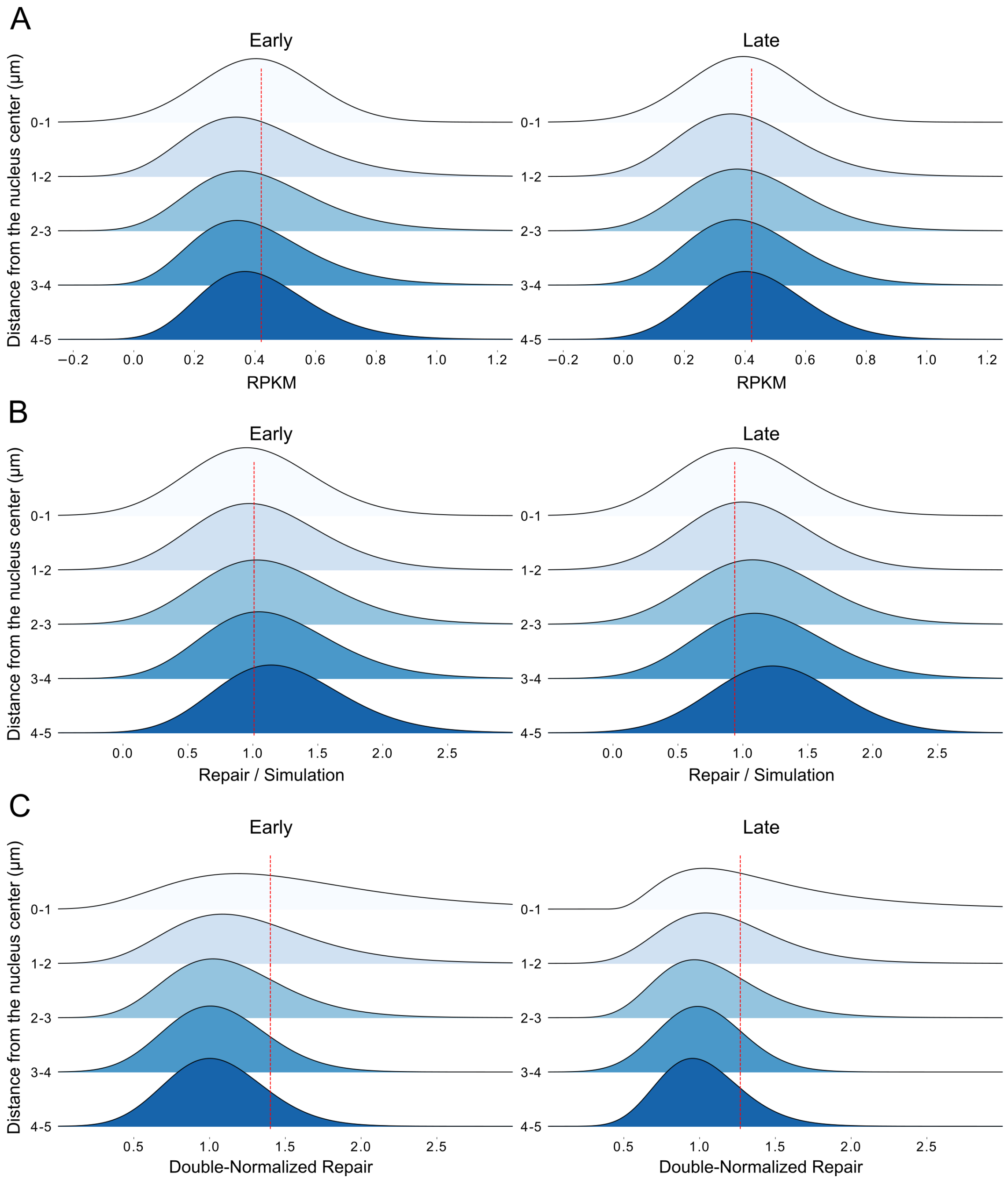

Figure A.3 Repair and normalized repair for CPDs in 3D layers.

Repair 2 hours after UV irradiation (A), Normalized (Repair / Simulation) (B) and Double-normalized (Damage-seq collected 2 hours after UV irradiation) (C) CPD early-phased (left) and late-phased (right) repair values on 1-micrometer genome slices. The density of repair values of the beads was shown. Dashed lines depict the median of the "0-1" region, described as a sphere with a 1-μm radius at the nucleus's center. Welch's t-test was performed to compare the “0-1” region against all the remaining regions (“1-2”, “2-3”, “3-4”, “4-5”), p-values for (A) are 0.84, 0.73, 0.97, 0.7 for early phased and 0.87, 0.53, 0.64, 0.25 for late phased respectively, p-values for (B) are 0.34, 0.021, 0.009, 0.0002 for early phased and 0.29, 0.009, 0.0019, 1.5e-05 for late phased respectively while p-values are (C) are 0.00014, 1.3e-07, 6.5e-10, 5.1e-10 for early phased and 0.0003, 3.4e-07, 4.26e-09, 4.33e-09 for late phased respectively.